EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Assessment of Interpolation Methods for EFAS

Giovanni Franchello

Institute for Environment and Sustainability

2005

EUR 22126 EN

European Commission
Directorate-General Joint Research Centre
Institute for Environment and Sustainability

Contact information
Address:Via E.Fermi 1, TP 670
E-mail: Giovanni.franchello@jrc.it
Tel.: ++39.0332.785066
Fax: ++39.0332.789007

http://natural-hazards.jrc.it/
http://www.jrc.cec.eu.int

EUR 22126 EN
Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

Preface

Acknowledgments

1. Data Sources
    1.1 MARS-STAT data base
        1.1.1 Grid data
        1.1.2 Station data
    1.2 High Resolution rainfall data in Central Europe
    1.3 Preliminary Comparisons
    1.4 Station data Processing

2. Interpolation methods
    2.1 Nearest and Inverse distance
        2.1.1 Calibration of the governing parameters
        2.1.2 Cross Validation
        2.1.3 Comparison of the Interpolated Data
    2.2 Fitting
        2.2.1 Cross Validation
        2.2.2 Comparison of the Interpolated Data
    2.3 Kriging
        2.3.1 Ordinary Kriging
        2.3.2 Cross Kriging
        2.3.3 Calibration of the governing parameters
        2.3.4 Cross Validation
        2.3.5 Comparison of the Interpolated Data

3. Filtering
    3.1 Calibration of the governing parameters
    3.2 Cross Validation
    3.3 Comparison of the Interpolated Data

4. Weighting
    4.1 Cross Validation
    4.2 Comparison of the Interpolated Data

5 - Cross Validation of the full set of the meteo variables

Summary and Conclusions

References

Appendix 1 – Equations for Data Analysis

Preface

The observed meteorological inputs for the European Flood Alert System (EFAS) have been analysed. Observed sources are: 1) Grid and Station data collected from the MARS-STAT data base, 2) high resolution rainfall and mean temperature data covering Germany and upper Danube (HR). Several interpolation methods have been calibrated and cross validated. The comparisons of the different interpolated data sources are discussed and the “best” interpolation method and parameters are suggested. The software station2map [3] is used to cross validate the different interpolation methods.

Acknowledgments

I thank the AgriFish Unit (JRC-IPCS), in particular Fabio Micale for providing the access and information’s on the MARS-STAT data base, Giorgio Liberta’ for giving the informatic’s support on accessing the data base, Janos Szabo for all the fruitful discussions on statistical analysis and developments in computer of the Kriging methods, Johan van der Knijff for providing useful bibliography on Geostatistics and Jutta Thielen for supporting such activity.

1. Data Sources

The data sources used for the assessment of the interpolation methods for EFAS are the JRC MARS-STAT data base, daily available from 1978 up to now, and the High Resolution Rainfall and Mean Temperature from Central Europe, available from year 1991 to 2002.

1.1 The MARS-STAT data base

The JRC-MARS project (Monitoring Agriculture through Remote Sensing techniques) provides timely information on the agricultural production to be expected in the current season, used by the European Commission for the implementation of the Common Agricultural Policy.

Available products include:

- maps of **weather indicators** based on observations and numerical weather models
- maps and time profiles of **crop indicators** based on agro-meteorological models
- maps of **vegetation indices** and **cumulated dry matter** based on remote sensing images

The weather observations and the numerical model results are included in a data base, organized by **Grid data** and **Station Data**, called **MARS-STAT data base**. Related information's are on the web site http://agrifish.jrc.it/marsstat/datadistribution/

The MARS-STAT Data Base contains data **from 1975**, covering the EU member states, the central European eastern countries, the new Independent states and the Mediterranean countries. The **DAILY** metereological data are the following:

maximum temperature (°C)
minimum temperature (°C)
mean daily vapour pressure (hPa)
mean daily windspeed at 10m (m/s)
mean daily rainfall (mm)
Penman potential evaporation from a free water surface (mm/day)
Penman potential evaporation from a moist bare soil surface (mm/day)
Penman potential transpiration from a crop canopy (mm/day)
daily global radiation in KJ/m^2/day
snow depth (cm) *   * data with no quality check

The Grid data (50x50 km grid size) are the results of a pre-interpolation process (CGMS) of an irregular network of weather stations (Fig. 1.1.1). The weather stations are not homogeneously distributed and not every MARS-STAT grid cell contains a weather station (Fig. 1.1.2).

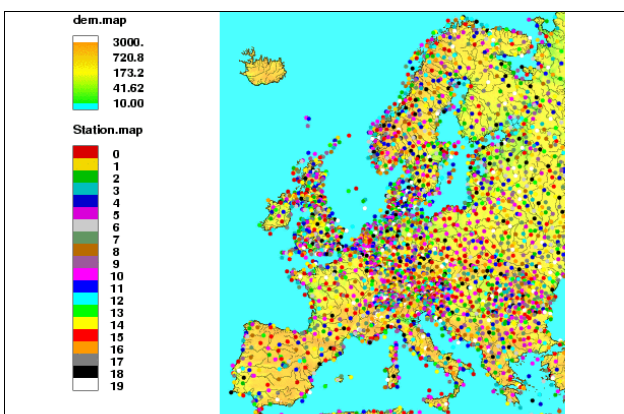


Fig. 1.1.1 – The MARS-STAT weather stations

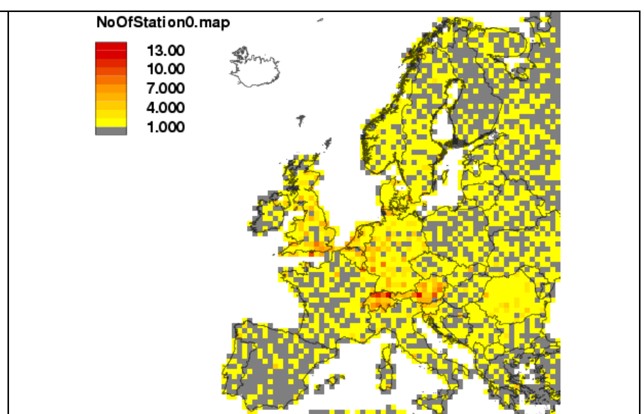


Fig. 1.1.2 – Number of stations per grid cell

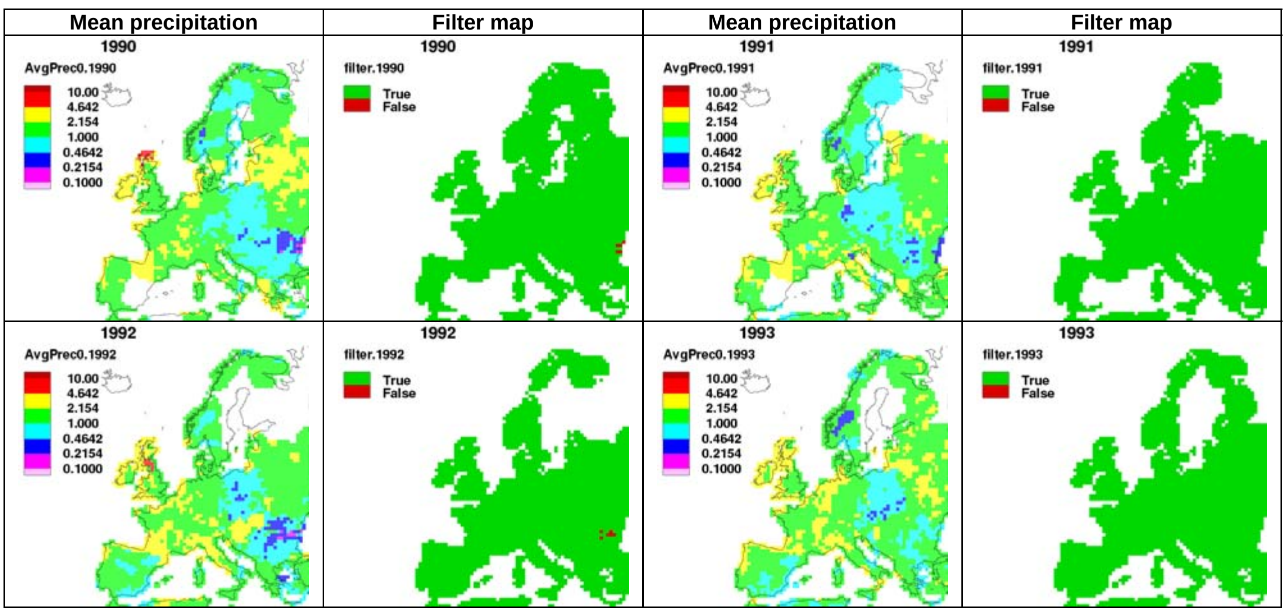
CGMS pre-interpolation [1] is performed in a way that the precipitation for a grid is taken from the weather station that is the most similar in position to the grid center (lowest score), while the other meteo variables are obtained by a weighted average based on scores. The above mentioned scores are obtained from the distance of the station to the grid point, the altitude difference, distance to coast and barrier blockage. In case the station data of a day is missing, CGMS will assign a value taken from a reference WEATHER-STATION table

1.1.2 Grid data

The precipitation data available in the MARS-STAT Grid data table (called here also Mars Grid data) has been analyzed from the year 1990 to 2004 (Fig. 1.1.3). One can realize the following:

- Years 1990 and 1991: the East of Spain is missing
- Years 1991 to 1994: data for most part of North Europe (Norway, Sweden, Finland, Lithuania, Estonia etc.) are missing
- Years 1990 and 1992: in Romania there are some pixel with very low (<0.4 mm/day) mean precipitation in the year
- Years 1990 to 1994: in Elbe and Danube there are large areas where the average precipitation is low compared with the nearest grid: after such period the values are consistent with the rest of Europe
- Years 1996, 1997, 2003, 2004: in Central Europe the are a few pixels with very low (<0.4 mm/day) mean annual precipitation.

At the moment, the policy taken by EFAS for pre processing the Mars grid data and make it available for the long period analysis (evaluation of thresholds for alert discharges) and water balance simulations (for setting up the initial condition of the hydrological forecasts) is to cover the missing data by inverse distance interpolation. The data are then interpolated into a 5x5 km grid, covering all Europe.



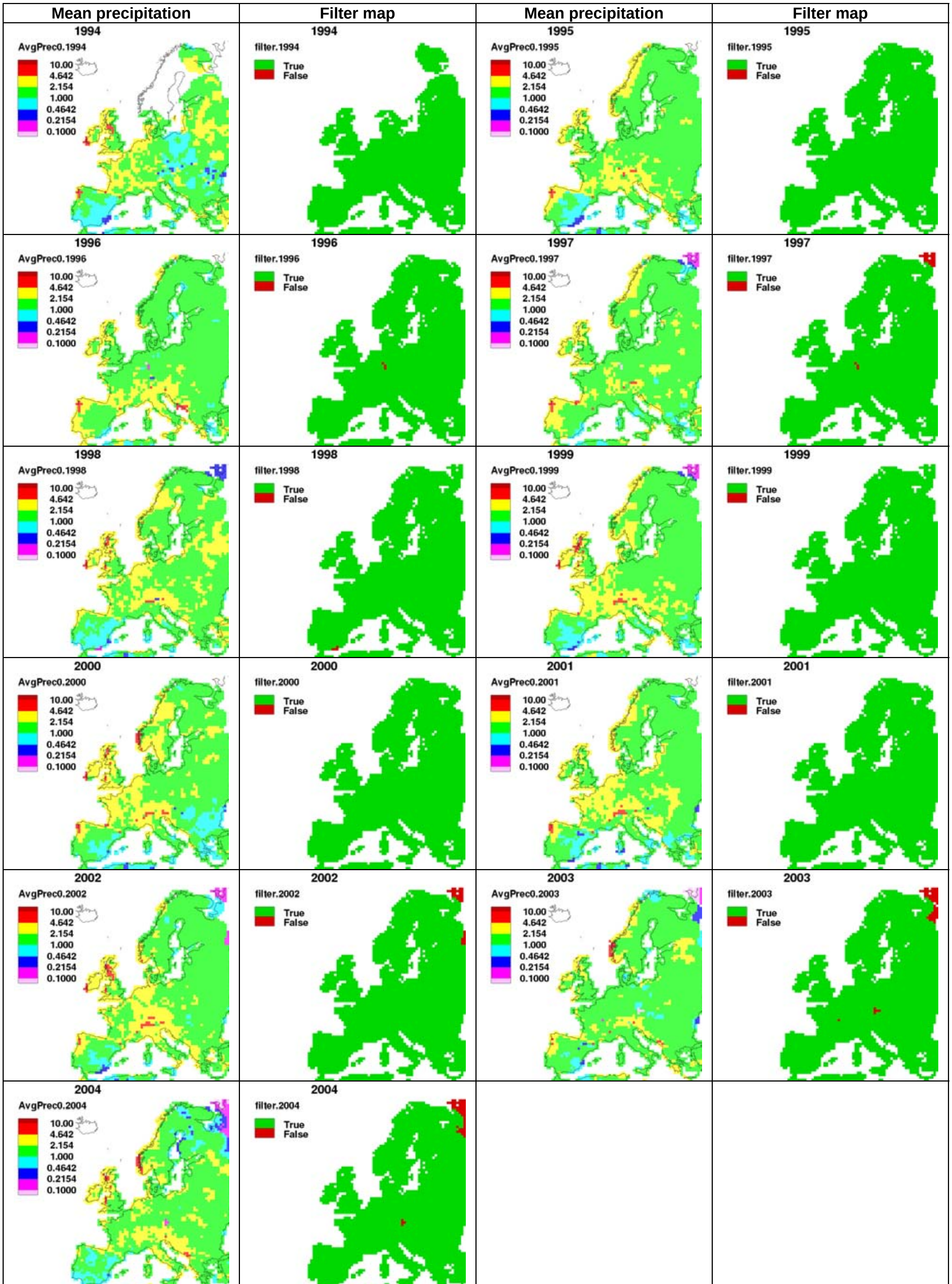


Fig. 1.1.3 – Annual mean daily precipitation and filter maps from MARS Grid data. In the filter maps with "False" are identified the pixels with mean daily rainfall below 0.4 mm/day

1.1.2 Station data

The weather station data stored in the MARS-STAT data base (called here also Mars Station data) are available from the year 1975 up to now. The no. of yearly available stations is shown in Fig. 1.1.4. The stations data base contain tables with the weather variables [1] labeled as Rain (i.e. precipitation), Temperature (i.e. minimum and maximum temperature) and Rest (calculated weather data, i.e. E0, ES0, ET0, Calculated radiation). In the data base, the station tables not necessarily contain the full set of the weather variables and the data are not necessarily available every day. In case the station data of a day is missing, CGMS will assign a value taken from a reference WEATHER-STATION table [1]. The no. of stations strongly increase after 1994.

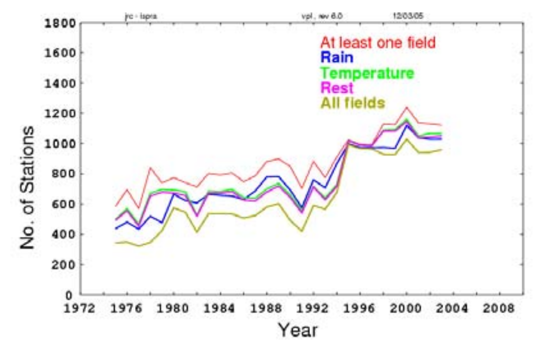


Fig. 1.1.4 – No. of Yearly Available Stations from WEATHER_STATION table

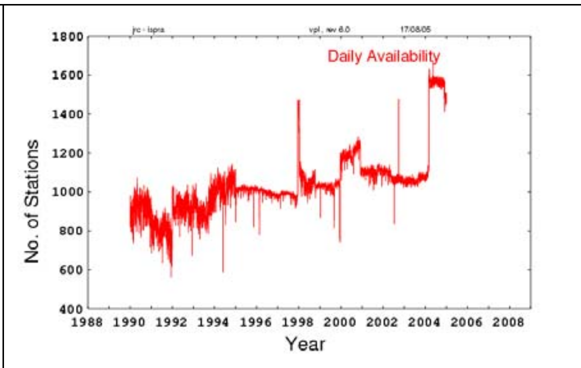


Fig. 1.1.5 – No. of Rain station daily availability

The no. of daily available Rain Stations, in the years 1994-2000, is shown in Fig. 1.1.5. One realizes that

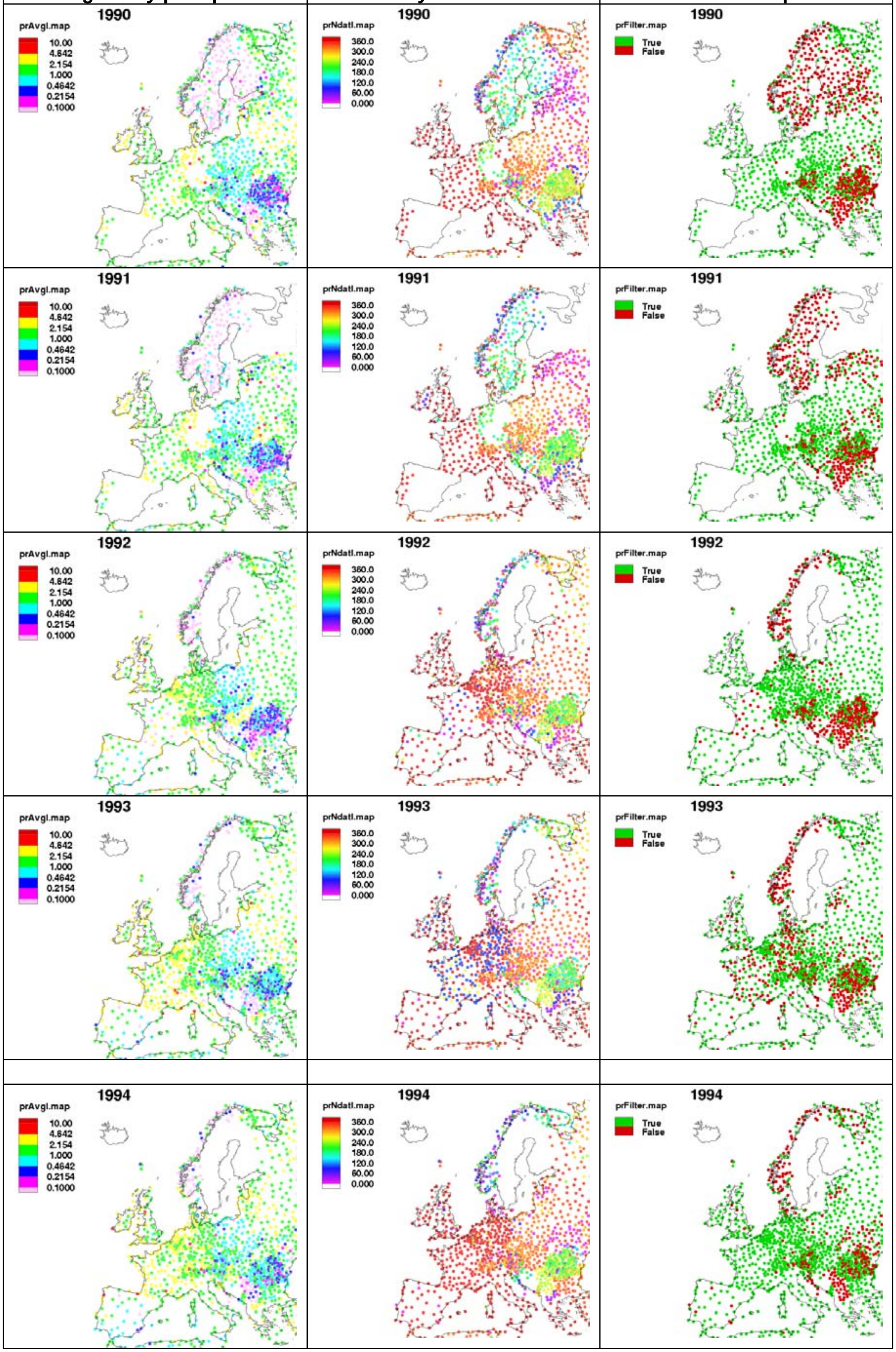
1) there are some "oscillations" in the no. of daily available stations
2) in the end of year 1997 to the beginning of year 1998 there is a peck in the no. of available stations
3) on year 2004 the no. of stations increase from ~ 1000 to ~ 1600

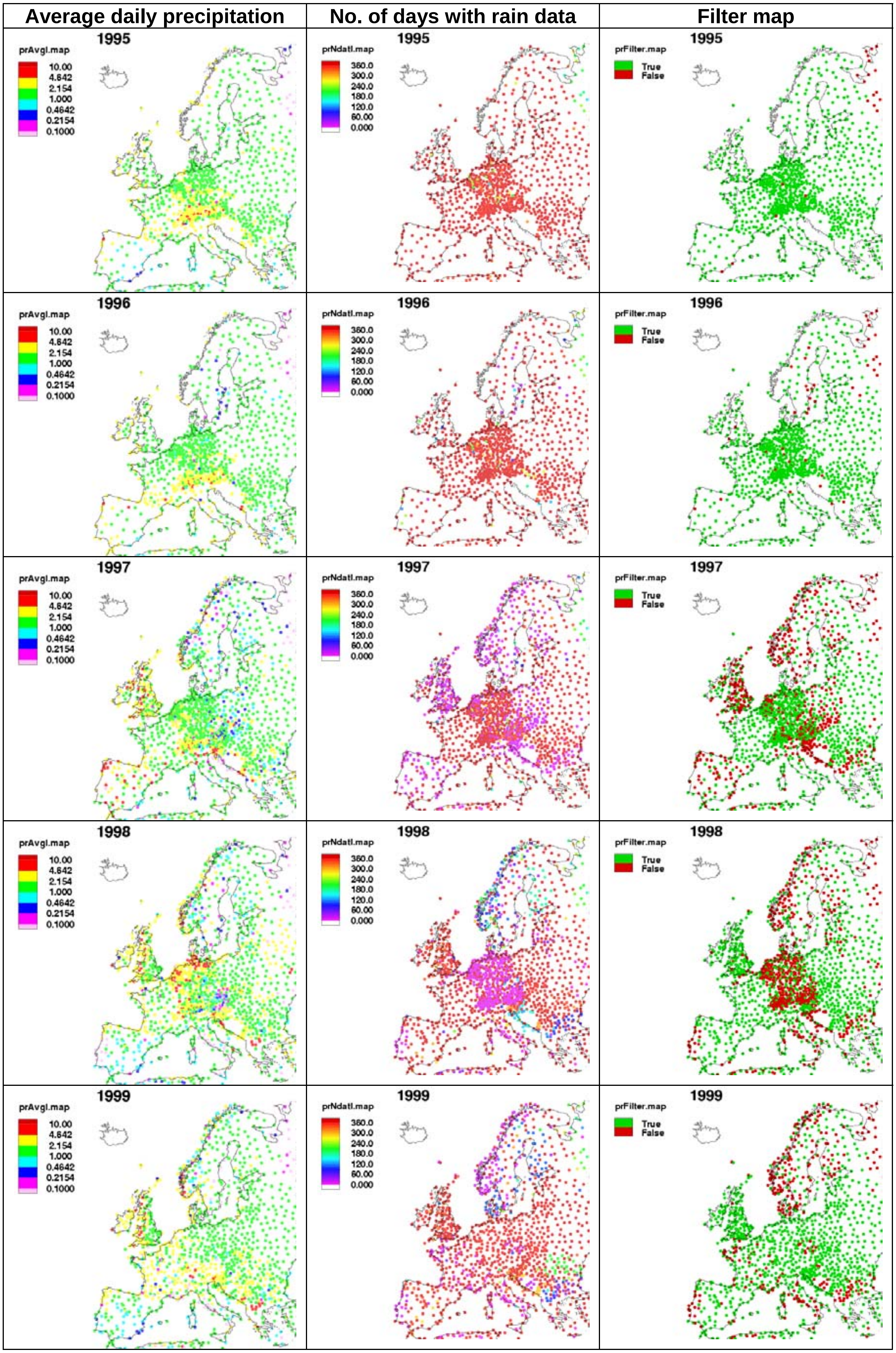
Looking in more detail the yearly space distribution of the rain Stations (Fig 1.1.6) is confirmed the lack of data in North Europe in 1994 (Norway, Sweden, Finland, Lithuania, Estonia etc.). In addition, there are some stations where the measured average rainfall is below 0.4 mm/day (~150 mm/year, considered a very low value for Europe, never observed) and the no. of days with rain data in the year is below 180, especially in 1994 and up to 1997. After the year 2000, the space distribution of the stations becomes more homogeneous, while the no. of stations in Italy, Spain and North Europe increase.

The yearly filter map, obtained by the following logical operator
Filter = (Mean daily rainfall > 0.4 mm/day) and (Mean daily rainfall < 20 mm/day)
and (no. of days with data available > 180)

allows to identify the stations that can have rain data with "suspected" values. The minimum and maximum annual mean daily rainfall have been chosen according to the historical observed rainfall in Europe. The no. of days with data available (180) is set lower of the threshold chosen by CGMS - which is 80 % of days in the year – in order to consider a larger no. of stations which can have "good" data.

Average daily precipitation	No. of days with rain data	Filter map




Average daily precipitation
No. of days with rain data
Filter map
1995
prAvgl.map
10.00
4.642
2.154
1.000
0.4642
0.2154
0.1000
1995
prNdatl.map
360.0
300.0
240.0
180.0
120.0
60.00
0.000
1995
prFilter.map
True
False
1996
prAvgl.map
10.00
4.642
2.154
1.000
0.4642
0.2154
0.1000
1996
prNdatl.map
360.0
300.0
240.0
180.0
120.0
60.00
0.000
1996
prFilter.map
True
False
1997
prAvgl.map
10.00
4.642
2.154
1.000
0.4642
0.2154
0.1000
1997
prNdatl.map
360.0
300.0
240.0
180.0
120.0
60.00
0.000
1997
prFilter.map
True
False
1998
prAvgl.map
10.00
4.642
2.154
1.000
0.4642
0.2154
0.1000
1998
prNdatl.map
360.0
300.0
240.0
180.0
120.0
60.00
0.000
1998
prFilter.map
True
False
1999
prAvgl.map
10.00
4.642
2.154
1.000
0.4642
0.2154
0.1000
1999
prNdatl.map
360.0
300.0
240.0
180.0
120.0
60.00
0.000
1999
prFilter.map
True
False

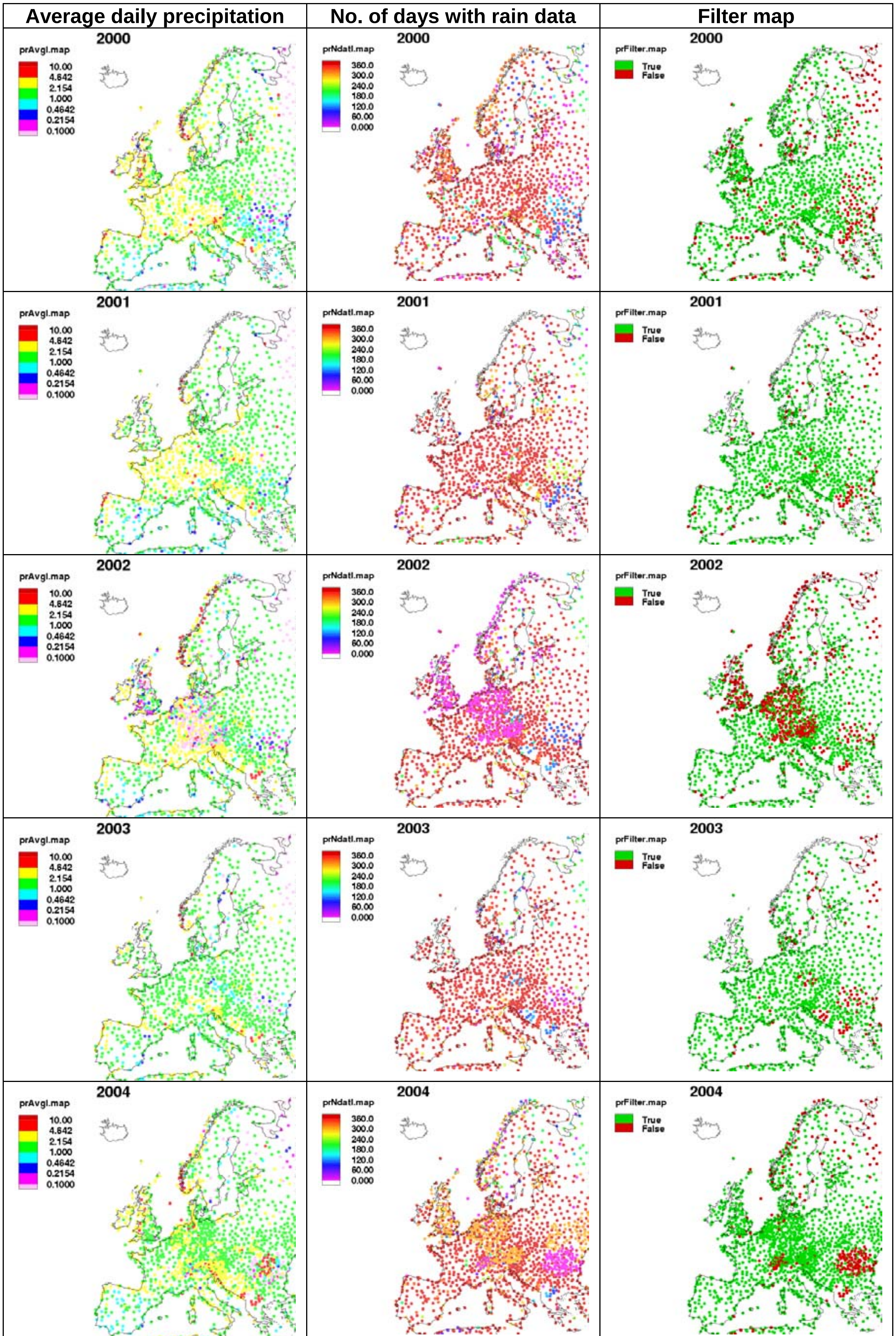


Fig. 1.1.6 - MARS Station data: yearly mean rainfall [mm/day], no. of days in the year with rain data and filter map

In Fig. 1.1.7 is the comparison between the no. of yearly available stations, obtained from the WEATHER_STATION table (red), and the no. of “filtered“ stations (blue), which are consistent: however, the filtering algorithm above explained seems less restrictive of the check algorithm performed by CGMS, which verify only that the percentile of days with data in a year is up to 80%.

After filtering the rain data, the no. of daily available stations is shown in Fig. 1.1.8: the no. of stations varies more smoothly during the years and the peck in the end of year 1997 disappears.

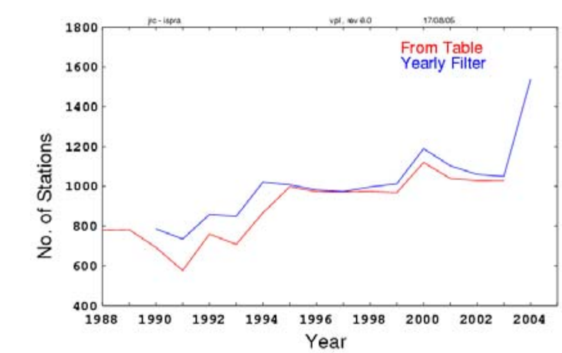


Fig. 1.1.7 – Yearly Station Availability

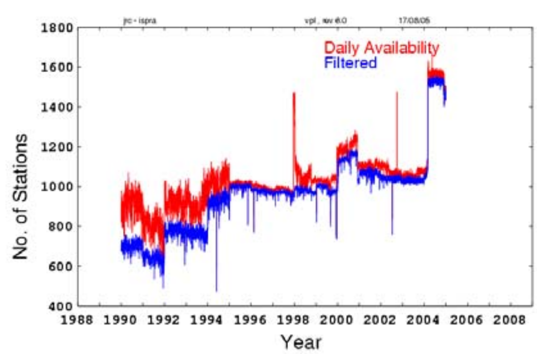


Fig.1.1.8 – Daily Station Availability

Applying the filter above described the following questions can arise

- the thresholds on minimum and maximum annual rainfall are set constant over all Europe, while they should be space dependent, i.e., such values should be different in Spain and Ireland.
- applying a too high threshold on the minimum days of data availability can lead to the result that “good data” are not used.

Therefore the thresholds should be based on a statistical analysis on the available data (see Section 4)

1.2 High Resolution rainfall data in Central Europe

A high resolution rainfall data set for Central Europe (Germany, Austria, Check Republic), covering the Elbe catchments and partially Rhine and Danube, is available for the period 1991 to 2002 (Fig. 1.2.1). The station availability per year is shown in Fig. 1.2.2: in the period 1994-2001 the area of Central Europe is covered by ~ 4000 Stations, while in the same area the Mars Stations are ~ 180: in this area each Mars grid cell (2500 km^2) is covered by ~ 15 High Resolution Stations and ~ 1 Mars Station.

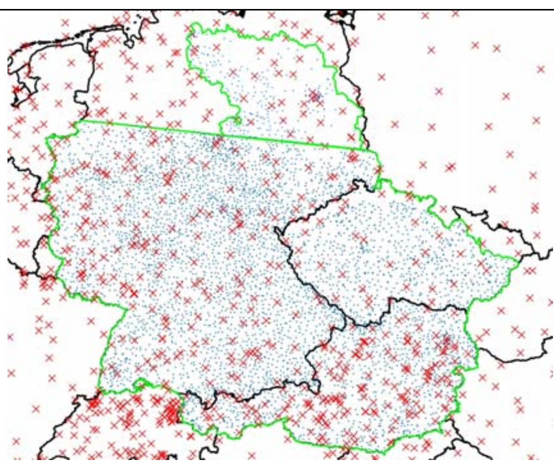


Fig. 1.2.1 – High resolution network (violet) compared with Mars station distribution (red)

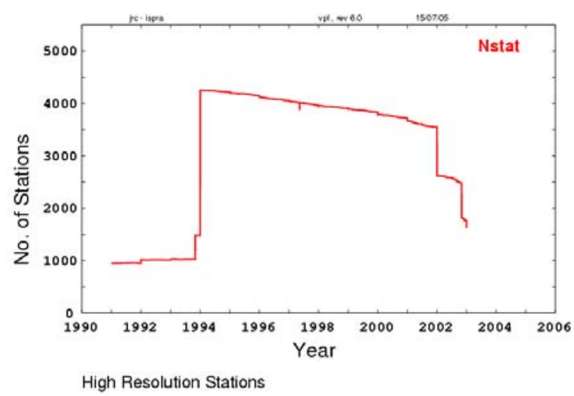


Fig. 1.2.2 – Daily available High Resolution Rainfall Stations

1.3 Preliminary Comparisons

A preliminary comparison of the MARS grid data with the HR rainfall data available for Central Europe is realized by calculating the Relative Root Mean Square Error (*RelRmse*) and the relative difference (*RelDiff*) described in Annex 1. More details on the data processing can be found in Section 1.4. After the comparison it can be seen that, after 1995, the MARS grid data underestimate the rainfall of ~10 % (Fig. 1.3.1 and 1.3.2), with a *RelRmse* of 60÷90 % (Fig. 1.3.3). Before 1995 the underestimation and the error were much larger. These differences can be imputed to the interpolation method adopted by CGMS to generate the 50x50 km grid data and/or to the quality of the Station data.

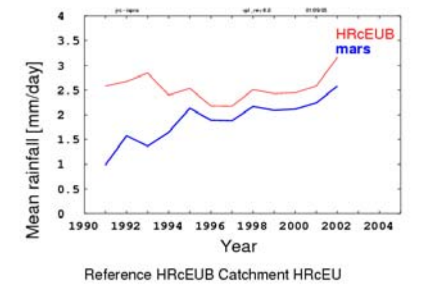
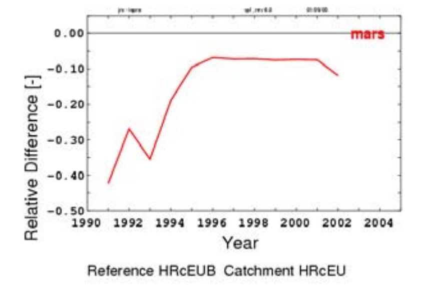
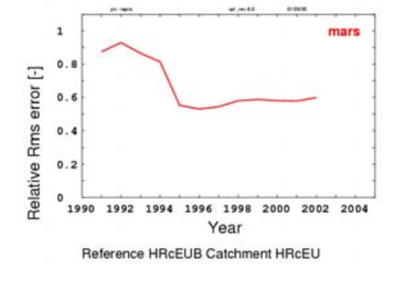
		

Fig. 1.3.1 – HR and MARS Grid mean annual rainfall	Fig. 1.3.2 – Relative difference from HR and MARS grid data	Fig. 1.3.3 – Relative Rms error of MARS grid data in respect to HR data

1.4 Station data processing

The Station data have been processed and interpolated by the following procedure:

1) the daily Station data are assigned to a 5x5 km grid: in case there are more than 1 station in a cell, the average is taken.
2) the 5x5 km cell with no station data (missing data) are interpolated looking to the nearest stations in a 200 km radius. The governing parameters of the different interpolation methods are calibrated and cross validated over the area shown in Fig. 1.4.1, covering all Europe (EuropeMars), the High Resolution area in Central Europe (HRcEU)
3) the data in the 5x5 km grid are averaged into a 50x50 km grid
4) The different data sets obtained from the MARS data have been compared with the HRcEU data.

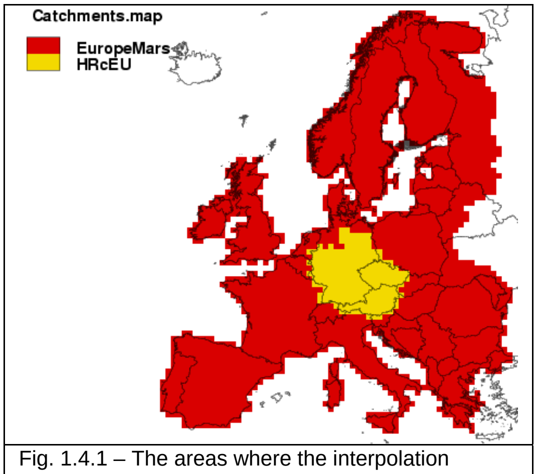


Fig. 1.4.1 – The areas where the interpolation methods are calibrated and cross validated

2. Interpolation methods

An assessment of different interpolation methods and data filtering of the available MARS Station data has been done in order to realize – if possible - an improvement of the input data for EFAS hydrological simulations. The strategy adopted by CGMS to pre-interpolate the Station data and generate the 50x50 Grid data has been developed for agriculture applications: however, for hydrological simulations is requested a higher resolution in space (and in time, as much as possible close to 1 hour resolution).

Therefore, the interpolation methods considered are the following

- Nearest
- Inverse distance
- Fitting
- Ordinal Kriging
- Cross Kriging

In addition are also analysed two statistical methods to filter and/or to weight the station data.

2.1 Nearest and inverse distance

The inverse distance interpolation method is well known and easy to apply: the governing parameters are the following (see Fig. 2.1.1)

- *Radius*: is the radius of the area where the nearest points are searched
- *Nmax*: is the maximum no. of points selected
- *Idp*: is the exponent of the inverse distance

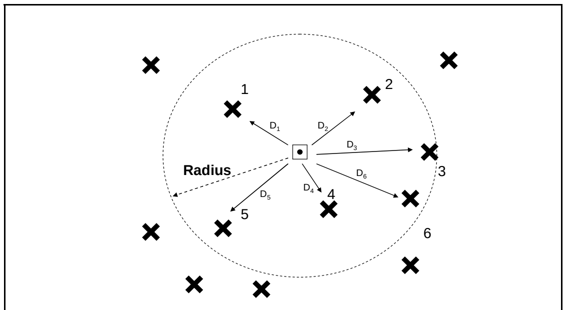


Fig 2.1.1 - the cross are the station positions while the square represents the cell where the value is interpolated

Each missing cell value *Z* is calculated by the following equations:

$$Z = \sum_{i=1,N\max} Z_i W_i \qquad \text{where} \qquad W_i = \frac{1/D_i^{\,Idp}}{\sum_{j=1,N\max} 1/D_j^{\,Idp}}$$

Defining *Nmax=1* the “nearest” method is applied.

2.1.1 Calibration of the governing parameters

The calibration of the parameters *Nmax* and *Idp* have been done for the years 1993 (poor data) and 1995 (good data) and verified in the areas in Fig. 1.3.1. Each station data is recalculated (source field) and compared with the original data (reference field) following the procedure described in Appendix 1.

Values of *Nmax* are 1,3,5,10,15,30 and *Idp*=1,2,3. For *Nmax*=1 the “nearest” interpolation method is applied. No filtering/weighting are applied, i.e., the original data are used.

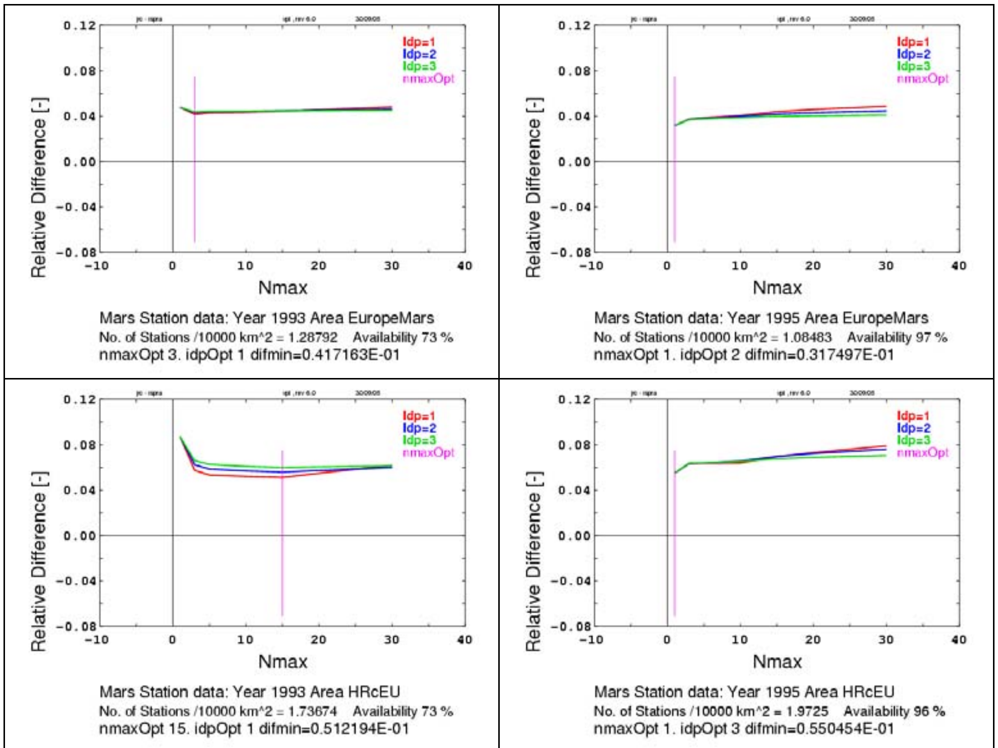


Fig 2.1.2 – *RelDiff* calculated for the years 1993 and 1995

The following behaviour are identified (see Fig. 2.1.2)

- The change of *RelDiff* in respect to *Nmax* and *Idp* is negligible for large areas
- Increasing *Idp* the change of *RelDiff* becomes lower
- The optimum combination of *Nmax*, *Idp* depends on the Year (or quality of the data) and on the Area under analysis; however, *Nmax* from 5 to 15 and *Idp*=2,3 are recommended

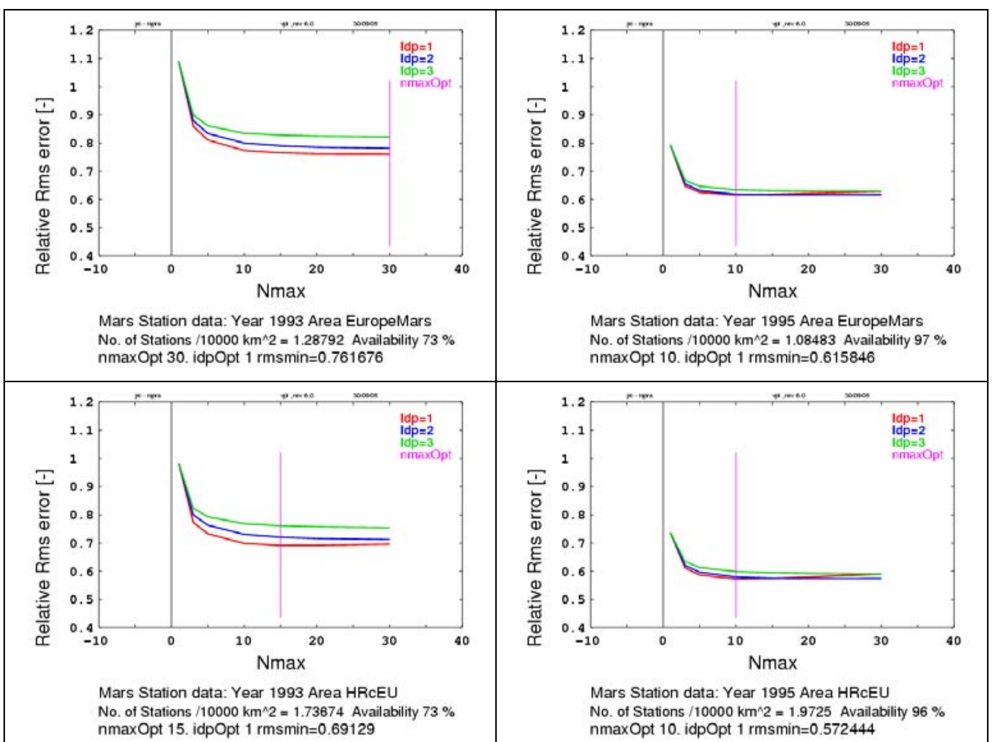


Fig 2.1.3 – *RelRmse* calculated for the years 1993 and 1995

The following behaviour are identified (see Fig. 2.1.3)

- The *RelRmse* reaches a minimum for *Nmax* > 10
- For *Nmax* > 15 the change of *RelRmse* becames small
- Increasing *Idp* the effect on *RelRmse* depends on the Year (or quality of the data): however, for large area and "good data" the effect of *Idp* becomes small comparing with *Nmax* effect
- The optimum combination of *Nmax*, *Idp* depends on the Year (or quality of the data) and on the Area under analysis; however, optimum *Nmax* is from 10 to 15, with *Idp*=1,2

2.1.2 Cross Validation

Combining the trends observed during the calibration process, 4 combinations of *Nmax, Idp* parameters have been identified (see Tab. 1)

Mars Data Set	HR data set	*Idp*	*Nmax*
marsStat0	HrStat0	-	1
marsStatA	HrStatA	1	15
marsStatB	HrStatB	2	10
marsStat3	HrStat3	3	5

Tab. 1 – The selected combination parameters for the Inverse Distance Method

The cross validation have been performed for the years 1990-2004 using the Mars Station data and the HR rainfall data in Central Europe, producing finally the graphs for each area under observation. No filtering is applied, i.e., the original data are used.

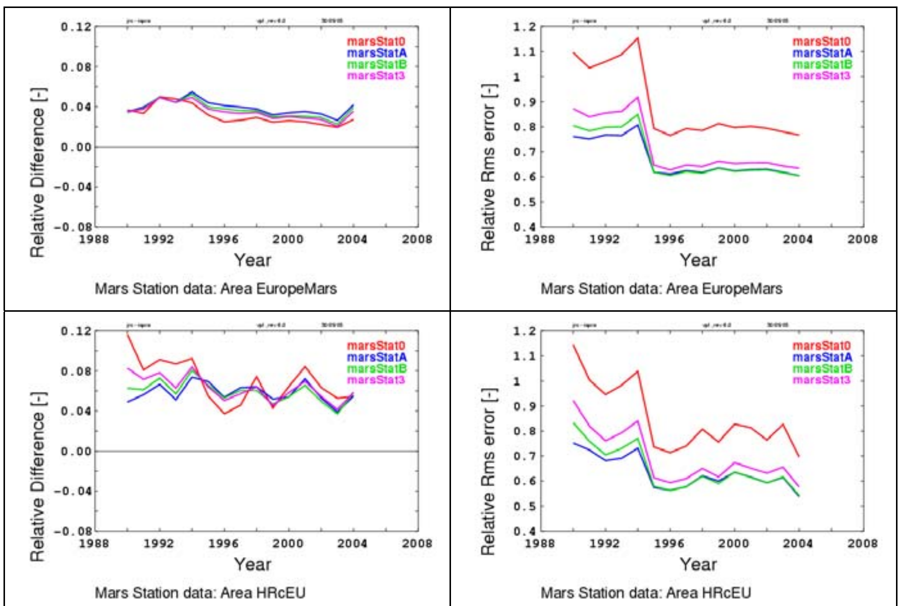


Fig 2.1.4 – Mars Station data plots: cross validation of the inverse distance method parameters

The cross validation on the Mars Station data leads to the following considerations (see Fig. 2.1.4 and 2.1.5)

- The *RelDiff* are in the order of ±4% while the *RelRmse* are in the order of 50÷100%
- The "nearest" method leads to a *RelRmse* higher of 20÷30% in respect to the other methods
- The *RelRmse* decrease after 1995, with values of ~ 80% for the nearest method and ~ 60 % for the other methods
- The *RelDiff* and *RelRmse* are higher in mountain regions and Mediterranean coasts
- The parameters used to generate the data set marsStatA and marsStatB are recommended.

RelDiff map | *RelRmse* map

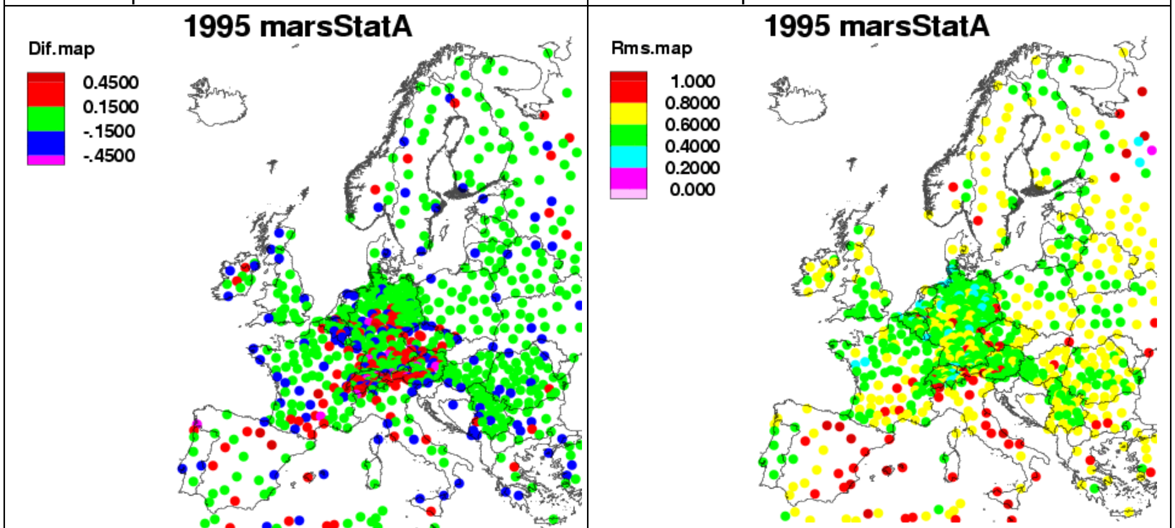


Fig 2.1.5 – Year 1995 Mars Station data maps: cross validation of the inverse distance method parameters.

The cross validation on the High Resolution rainfall data lead to the following considerations (see Fig. 2.1.6)

- The *RelDiff* are ~ 1% while the *RelRmse* are in the order of 35%
- The "nearest" method leads to a *RelRmse* higher of ~10% in respect to the other methods
- The *RelRmse* is constant for the period under analysis, with values of 43÷46% for the nearest method and 33÷37 % for the other methods
- The parameters used to generate the data set HrStatA are recommended (lower *RelRmse* and negligible differences for the *RelDiff*)

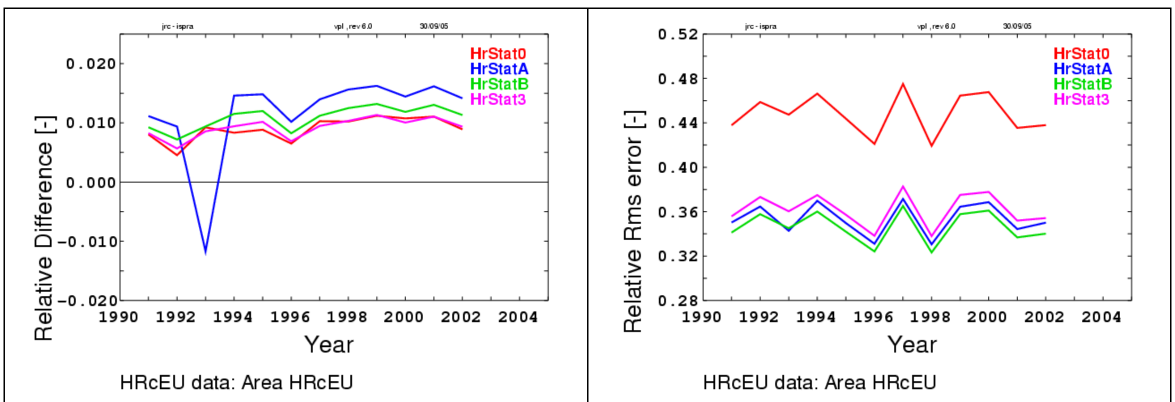


Fig 2.1.6 – High resolution rainfall data plots: cross validation of the inverse distance method parameters

2.1.3 Comparison of the Interpolated Data

The HR rainfall data available for Central Europe have been interpolated onto a 5x5 km grid setting the parameters *Nmax*=10 and *Idp*=2 (HrStatB data set) and then rescaled to a 50x50 km grid. The same procedure is repeated to the different data sets obtained from the mars Station data and applying the parameters as described in Tab. 1. The data set labeled “mars” stand for Mars Grid Data. The mean annul rainfall, the *RelDiff* and the *RelRmse* quantities are calculated, taking as reference the HR data set. The mean annual rainfall is still underestimated (Fig 2.1.7) but with a small improvement in respect to the Mars Grid data, maintaining the same behavior for each data sets. A remarkable improvement is seen (Fig 2.1.8) for the *RelRmse* of the mars Station data sets, which is decreased of ~20% in respect to the mars Grid Data set.

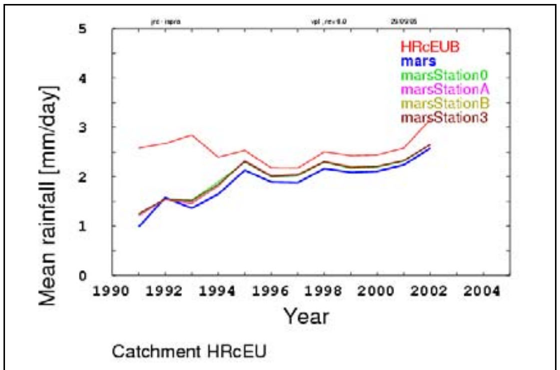


Fig 2.1.7 – Mean annual rainfall: see Tab 1 for the legend explanation. The mars Gid data are labeled as “mars”

No big differences are seen between the mars Station data sets: however, marsStationA and marsStationB data sets seem the best option on reproducing the HR data

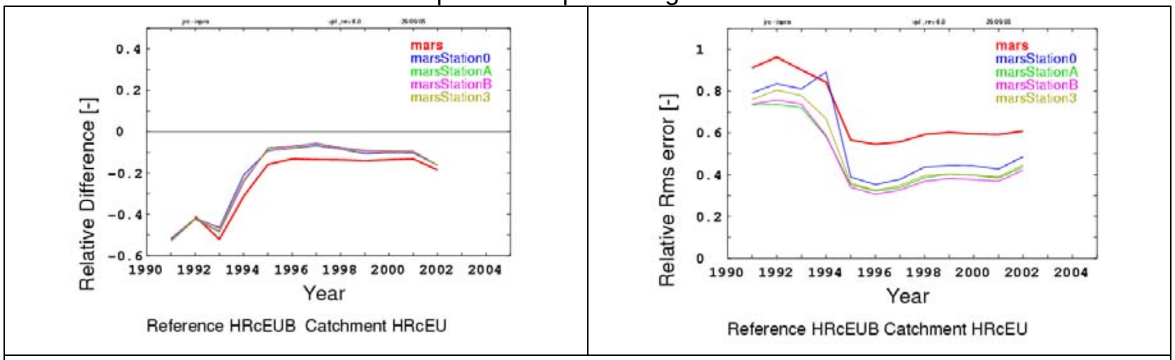


Fig 2.1.8 – Comparisons of mars Grid data (mars) and mars Station data (marsStation0,A,B,3) with the HR data set in Central Europe

In Fig. 2.1.12 are shown the rainfall maps related to the floods in Danube in August 2002: such maps are included with the scope to show the rainfall pattern realised with the different parameter combinations. The no. of selected stations (*Nmax*) strongly influences the rainfall patterns. If *Nmax* increases, the surface becomes more smooth and more “nice” to see, but with the danger to loose information on the “real structure” of the rainfall patterns, which is in any case impossible to capture given the limited no. of stations (Fig. 2.1.9) comparing to the High Resolution rainfall data (Fig. 2.1.10)

In Fig. 2.1.13 are shown the rainfall maps related to the floods in Switzerland, Austria and Germany in the period 21-23 August 2005: the behaviours of the rainfall patterns is similar to year 2002, but less strong because the space resolution of the stations increase significantly (Fig 2.1.11). Note from the figures shown that on August 2002 the rainfall intensity was about the double comparing with August 2005.

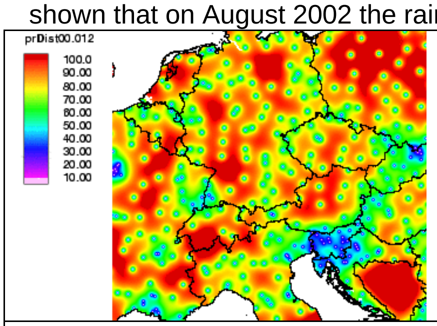


Fig 2.1.9 – Distance [km] of the interpolated grid values to the Mars Stations on 12 August 2002

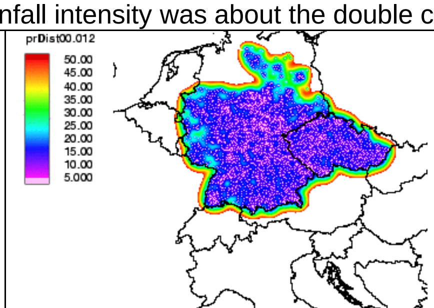


Fig 2.1.10 - Distance [km] of the interpolated grid values to the HR Stations on 12 August 2002

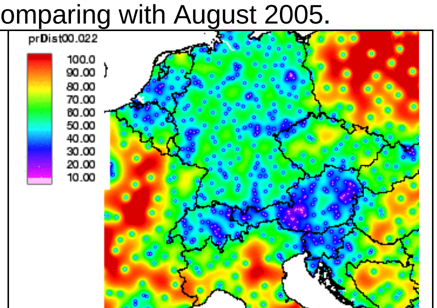


Fig 2.1.11 - Distance [km] of the interpolated grid values to the Mars Stations on 22 August 2005

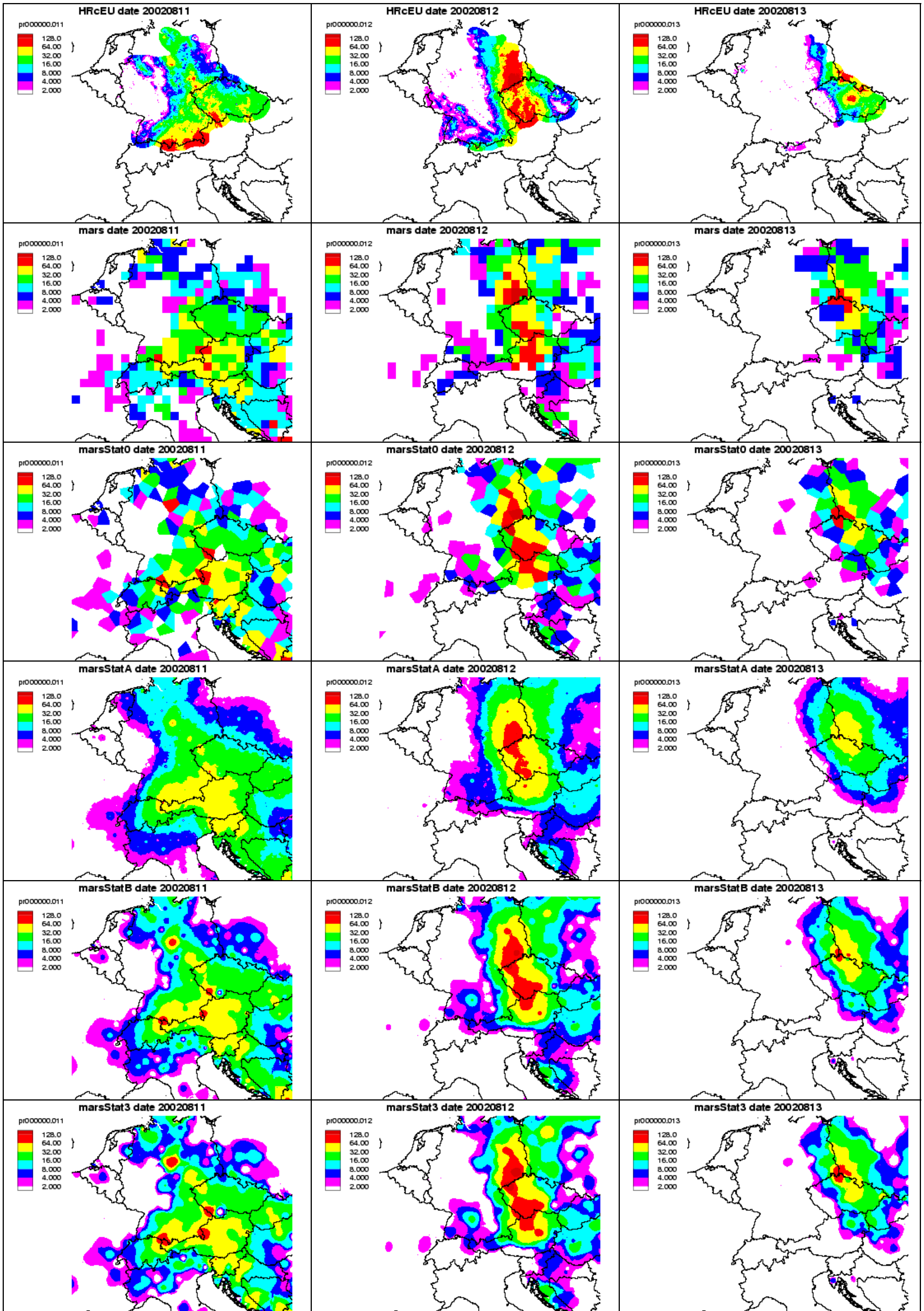


Fig. 2.1.12 – Rainfall maps for the period 11-13 August 2002

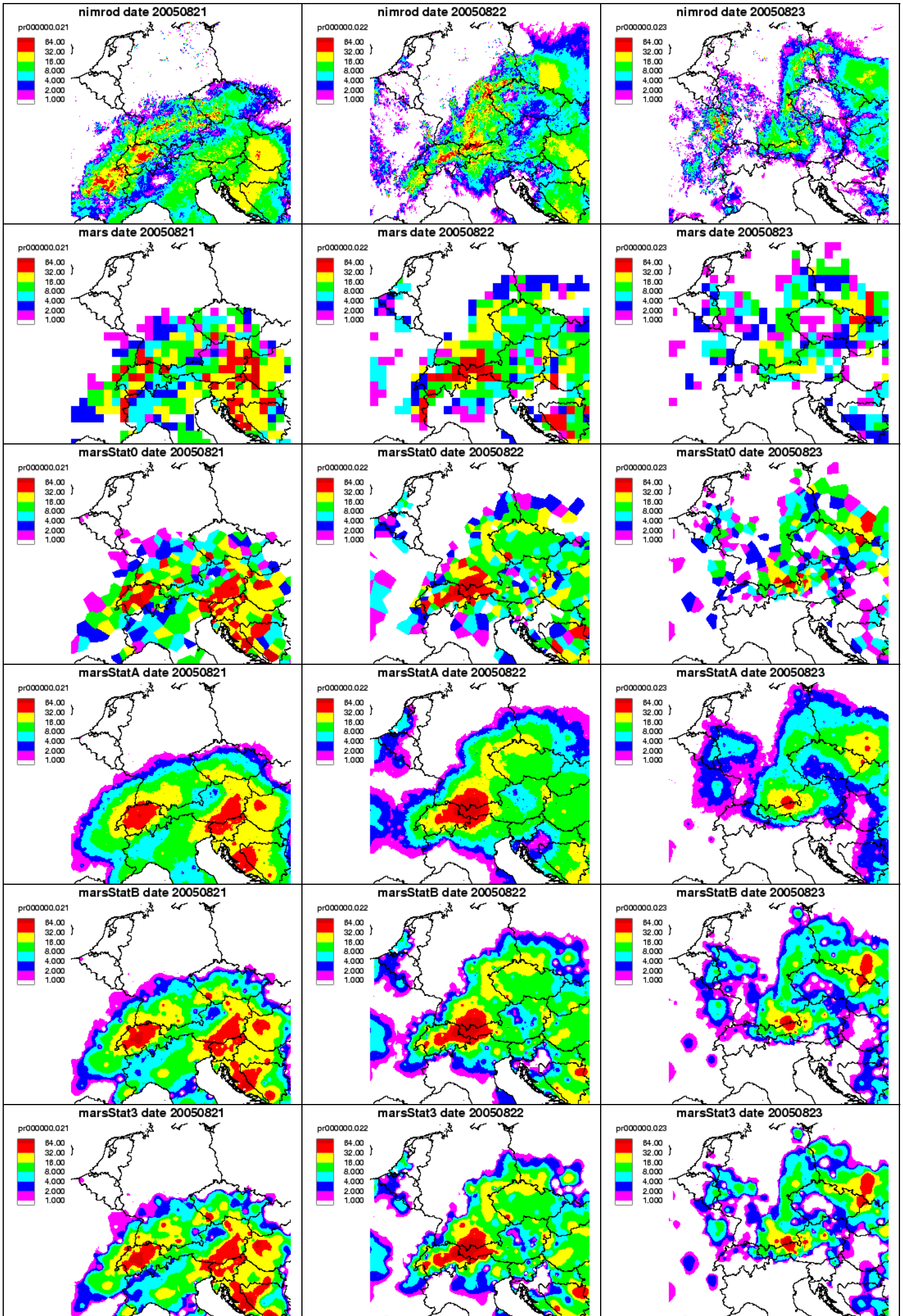


Fig. 2.1.13 – Rainfall maps for the period 21-23 August 2005: in the top of the figure, the maps with the title "nimrod" are generated by using radar data provided by UK MetOffice Nimrod system

2.2 Fitting

Introducing the fitting method is a first attempt in the direction of "filtering" or smoothing the suspicious values. The fitting method consists in evaluating for each point to be interpolated the surface $Z(X,Y)$ that **best fit** the *Nmax* nearest points in the area identified by *Radius* (see Fig.2.2.1)

- *Radius*: is the radius of the area where the nearest points are searched
- *Nmax*: is the maximum no. of points selected
- *Idp*: is the exponent of the inverse distance
- D_i is the distance of the stations to the point to interpolate
- *Z* is the value to interpolate (rainfall, temperature etc)

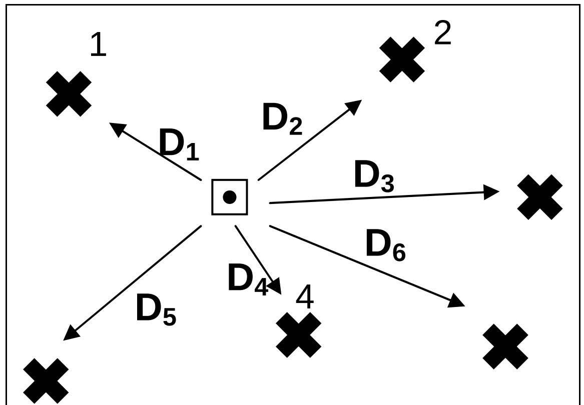


Fig 2.2.1 - the cross are the station positions while the square represents the cell where the value is interpolated

To each point is given the following weight: $W_i = \dfrac{1/D_i^{Idp}}{\sum_{j=1,N\max} 1/D_j^{Idp}}$

The surface $Z(X,Y)$ can be of the first or second order

Order 1: $Z(X,Y) = a + bX + cY$

Order 2: $Z(X,Y) = a + bX + cY + dX^2 + eY^2 + fXY$

Where X and Y are the coordinates of the point to interpolate and the coefficients *a,b,c,d,e,f* are evaluated by a weighted multiple regression based on X_i, Y_i, Z_i, W_i of each Stations in the *Radius* area. Note that the Inverse Distance Method is in reality a weighted multiple regression of a surface of order 0, i.e., a horizontal plane. The fitting method allows evaluating the space derivatives of the $Z(X,Y)$ value, which is not possible with the other methods.

2.2.1 Cross Validation

Cross validation of the fitting method of order 1 and 2 have done setting *Nmax=15,* high enough in order to have a sufficient no. of points for the multiple regression (see Tab. 2)

Mars Data Set	HR data set	*Idp*	*Nmax*	Method	Order
marsStatA	-	1	15	0	0
-	HrStatB	2	10	0	0
marsStatF1	HrStatF1	1	15	1	1
marsStatF2	HrStatF2	1	15	1	2

Tab. 2 – The selected parameters for the Fit Method

No pre-filtering is applied, i.e., the original Mars Station data and the HR rainfall data in Central Europe are used.

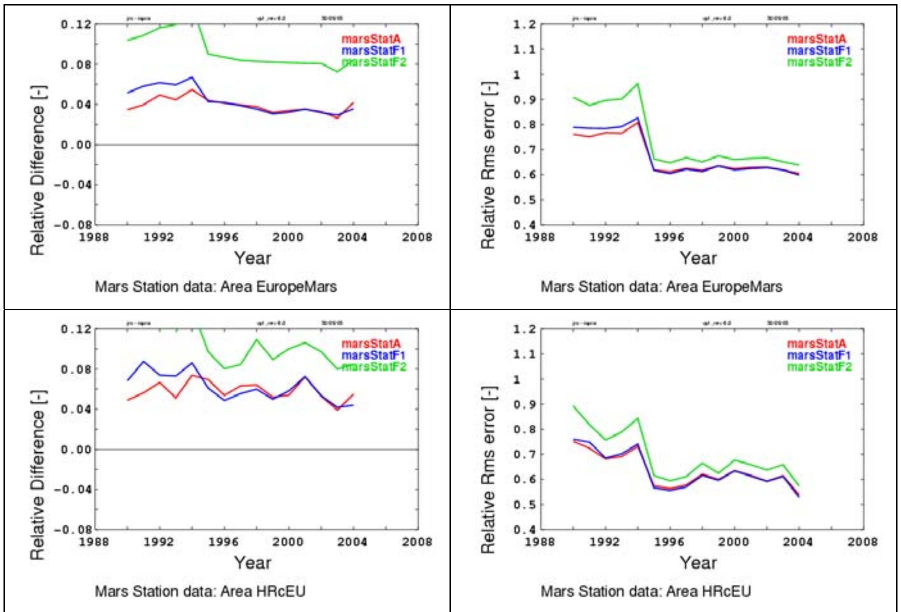


Fig 2.2.2 – Mars Station data plots: cross validation of the fitting methods

In Fig 2.2.2 and 2.2.3 is shown that fitting the Station data with a surface of order 1 the results are similar with the Inverse distance method, while with a surface of order 2 the *RelRmse* and *RelDiff* quantities increase.

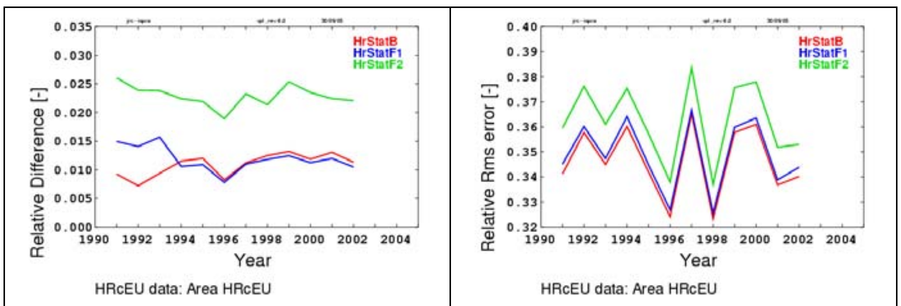


Fig 2.2.3 – High resolution rainfall data plots: cross validation of the fitting methods

2.2.2 Comparison of the Interpolated Data

The Mars and HR Station data are processed following the same procedure described in section 2.1.3. After a comparison of the mars Station data with the HR data (Fig. 2.2.4) is confirmed what seen in the cross validation: the fitting of order 1 gives similar results of the Inverse Distance Method, while for fitting of order 2 the results are worst.

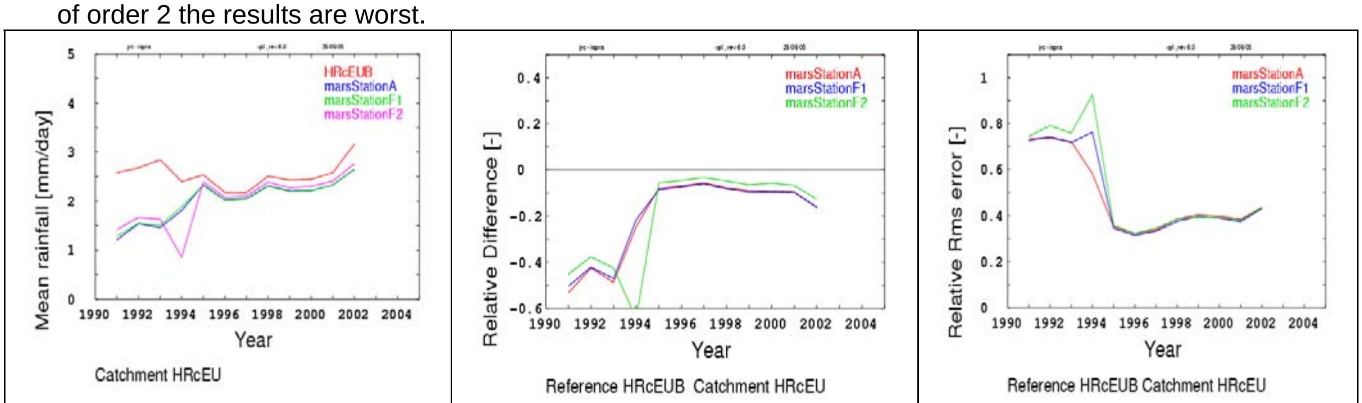


Fig 2.2.4 – Comparisons of mars Station data (marsStationA,F1,F2) with the HR data set in Central Europe

2.3 Kriging

Geostatistical methods of interpolation, popularly known as *Kriging*, differ from the inverse distance method by the way the interpolation weights are evaluated: for Kriging, such weights are evaluated by a statistical analysis and regression of the spatial variation of the data, distinguish between a) deterministic variation of the data – physically easy to explain, like altitude variation etc. -, b) spatially auto correlated variation, c) uncorrelated noise. [2]. Such statistical method allows to estimate the standard error associated to the interpolated values.

2.3.1 Ordinary Kriging

- *Radius*: is the radius of the area where the nearest points are searched. The experimental variograms are built using the points *N* found in the area
- *Nmax*: is the maximum no. of points selected to evaluate the weights
- D_{ij} is the distance from 2 stations
- M: no. of pairs $V_{i,j}, D_{i,j}$ obtained from the points *N* found in the area

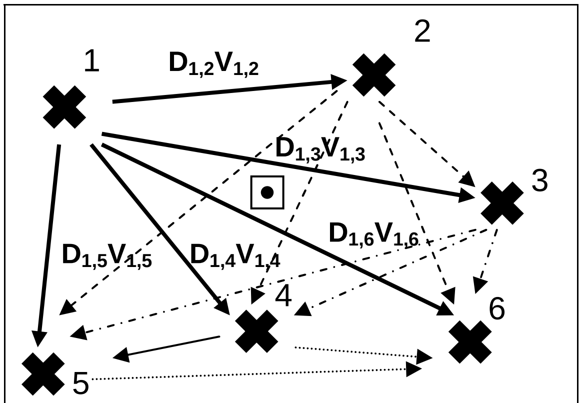


Fig 2.3.1 – The cross are the station positions while the square represents the cell where the value is interpolated. In the figure, for N=6, the no of pairs $V_{i,j}, D_{i,j}$ are $\mathrm{M} = \mathrm{N}(\mathrm{N}-1)/2 = 15$

Briefly, the scheme followed to compute the interpolation weights is the following

1) Evaluation of the variograms (see Fig 2.3.1)
   a. Build the experimental variogram by computing the *M* pairs $V_{i,j}, D_{i,j}$ found in the area identified by *Radius,* with $V_{ij} = 0.5(Z_i - Z_j)^2$ and $D_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}$.
   b. Evaluate the variogram function *varFun(Lag(D))=varFun(D)* by fitting the pairs V_{ij} and $Lag(D_{ij})$, where the Lag function is one of the variogram models - described below - that best fit the experimental variogram. The search of the best variogram model is done automatically. The variogram function will be

$$VarFun(D) = \mathrm{var}\,Fun(Lag(D)) = c_0 + c_1 Lag(D)$$

   where c_0 is called "nugget". **Note that the automatic fitting can lead to negative coefficients:** in this case is imposed $c_0 = 0$ and the fit is repeated again.
   c. Ignore the data points where the difference between calculated variance and experimental variograms is higher then a threshold (which is calibration parameter) i.e.,

$$abs\left[V_{i,j} - \mathrm{var}\,Fun\left(D_{i,j}\right)\right] > Rmse\ \ CleanFactor$$

where

$$Rmse = \sqrt{\frac{\sum_{i=1,n-1}\left[\sum_{j=2,n}\left(V_{i,j} - \mathrm{var}\,Fun(D_{i,j})\right)^2\right]}{\mathrm{M}}}$$

2) Calculate the interpolation weights by solving the linear equation of *Nmax+1* unknown

$$A \cdot \begin{bmatrix} W \\ \phi \end{bmatrix} = b$$

where

$$A = \begin{bmatrix} VarFun(D_{1,1}) & \cdots\cdots & VarFun(D_{1,N\max}) & 1 \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & 1 \\ VarFun(D_{N\max,1}) & \cdots\cdots & VarFun(D_{N\max,N\max}) & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}; \qquad b = \begin{bmatrix} VarFun(D_1) \\ \cdots\cdots \\ VarFun(D_{N\max}) \\ 1 \end{bmatrix}$$

and W_i are the interpolation weights and ϕ is a Lagrangian to solve the linear system. **Note that the solution of the system equation can lead to negative weights:** after having explored different strategies – a) repeat the calculation excluding the fares points until all weights are positive b) exclude the points where the calculated weight are negative – the solution adopted is simply to accept only the positive weights (and dived each weights by the sum of the weights).

3) Calculation of the interpolated data *Z* and the associated standard error σ.

$$Z = \sum_{i=1,N\max} Z_i W_i \qquad\qquad \sigma = \sqrt{\phi + \sum_{i=1,N\max} \operatorname{var} Fun(D_i) W_i}$$

The variogram models considered are the following (Fig. 2.3.2)

1 – Linear: Lag(h) = h

2 – Spherical:

$$Lag(h) = \begin{cases} 1.5h - 0.5h^3 & if \quad h < 1 \\ 1 & if \quad h \geq 1 \end{cases}$$

3 – Exponential: $Lag(h) = 1 - \exp(-1.5h)$

4 – Quasi- exponential:

$$Lag(h) = 1.5h/(1+1.5h)$$

5 – Gaussian: $Lag(h) = 1 - [\exp(-1.5h)]^2$

11–Logarithmic:

$$Lag(h) = \ln(D)/\ln(Range)$$

The ratio h is multiplied by 1.5 (Exponential, Quasi exponential and Gaussian models) in order that the derivative $\partial Lag(h)/\partial h = 1.5$ for h=0 as for the spherical model.

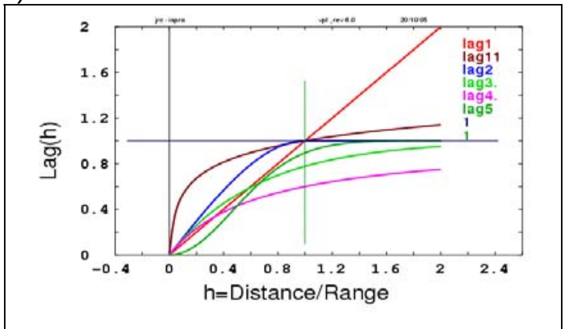


Fig. 2.3.2 – Variogram models as function of h=Distance/Range, where Range is 95% of the distance where the variance reach the maximum. In the legend, lagN identify the variogram model N, i.e., lag3 identify the exponential model.

In Fig. 2.3.3 and 2.3.4 are shown the variogram models “exponential to linear with sill” and “quasi exponential to linear with sill”. They are a combination of the linear with sill model with the exponential or “quasi exponential” models. Note that the “quasi exponential” variogram model increase less faster comparing with the exponential one.

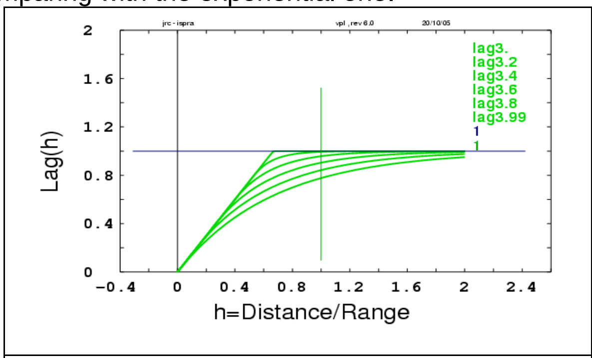


Fig 2.3.3 – Exponential (lag3.) to linear with sill (lag3.99) models.

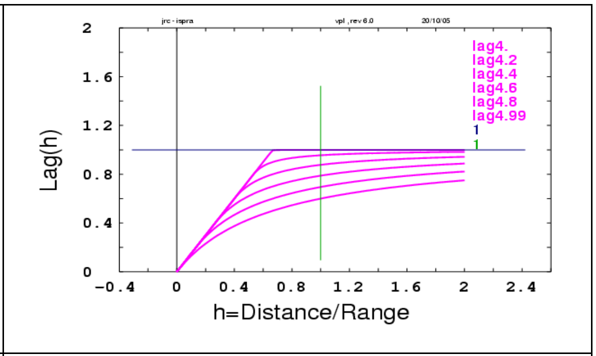


Fig 2.3.4 – Quasi Exponential (lag4.) to linear with sill (lag4.99) models.

The equation of “exponential to linear with sill” is the following

$$Lag(h,r) = \begin{cases} r & if \quad 1.5h \le r \\ r + d1 * [1 - \exp(-d2/d1)] & if \quad 1.5h > r \end{cases}$$

where *r* is a ratio that can vary from 0 (exponential model) to ~1 (linear with sill model).

The equation of “quasi-exponential to linear with sill” is the following

$$Lag(h,r) = \begin{cases} r & if \quad 1.5h \le r \\ r + d1 * d2/(d1 + d2) & if \quad 1.5h > r \end{cases}$$

where *r* is a ratio that can vary from 0 (quasi-exponential model) to ~1 (linear with sill model).

In both models $d1 = 1 - r \quad and \quad d2 = 1.5h - r$

2.3.2 Cross Kriging

- *Radius*: is the radius of the area where the nearest points are searched. The experimental variograms are built using the *N* points found in the area
- *Nmax*: is the maximum no. of points selected to evaluate the weights
- D_i: mean distance from the station *i* to the nearest Stations in the area idenfied by *Radius*
- M=N: no. of pairs V_i, D_i

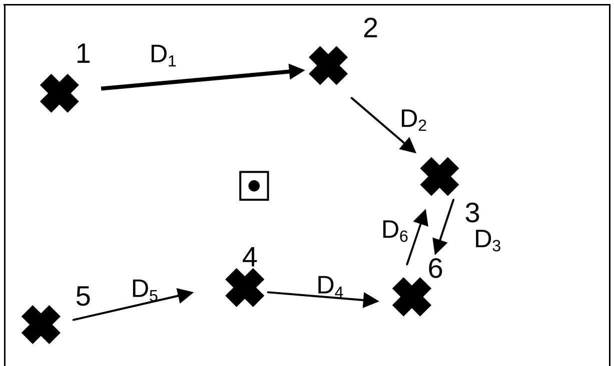


Fig 2.3.5 - The cross are the station positions while the square represents the cell where the value is interpolated. In the figure, for N=6, the no of pairs V_i, D_i are $\mathrm{M} = 6$

The Cross Kriging method introduced in this work vary from the Ordinary Kriging method by the way the experimental variograms are evaluated (see in section 2.3.1, step 1.a). In this case (Fig. 2.3.5) the experimental variograms are built by the pairs V_i, D_i, with $V_i = 0.5(Z_i - Z_{i,rec})^2$ and $Z_{i,rec}$ evaluated by a cross validation of the station data fields. The no. of pairs corresponds to the no. of points in the Radius area, lower of the *M* pairs found in the Ordinary Kriging method. The other steps (section 2.3.1, step 1b, 1c, 2,3) are the same as for the Ordinary Kriging. **Problems can arise if the stations are homogenously distributed:** in such a case all the distance will be the same, leading to problems on fitting the experimental variograms (negative nugget or error from the fitting solver): in this case the nugget is set to zero and the fitting is repeated agin. A higher *Radius* can overcome the problem. The cross Kriging method is introduced as an attempt to minimise the error evaluated by the cross validation process.

2.3.3 Calibration of the governing parameters

The calibration has been done for the years 1993 (poor data) and 1995 (good data). Each station data is recalculated (source field) and compared with the original data (reference field) following the procedure described in Appendix 1. A first calibration has done in respect to the values *Nmax*=5, 10, 15, 20, 30 and without cleaning (step 1c not done). The variogram models to select are set to typlag=1, 2, 3, 3.4, 3.8, 4, 4.4, 4.8, 5, 11 (see Figs. 2.3.2-2.3.4 about the variogram models).

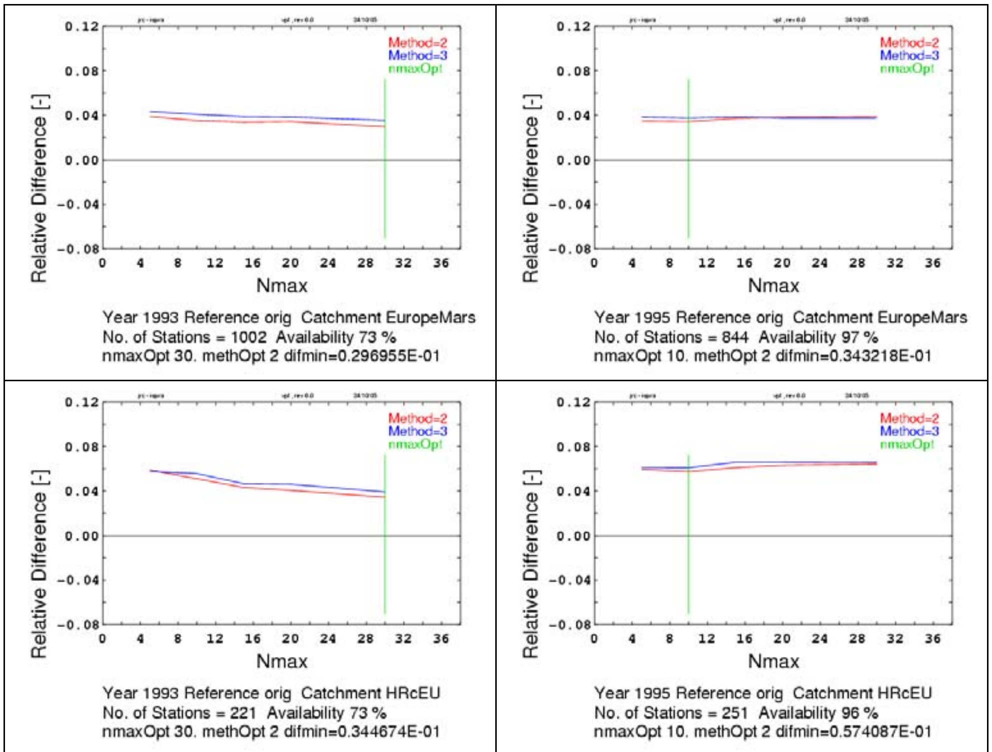


Fig 2.3.6 – Relative Difference calculated for the years 1993 and 1995: Method=2 stands for Ordinary Kriging while Method=3 stands for Cross Kriging

Fig. 2.3.6 shows that the quantities *RelDiff* and RelRmse are quite insensitive in respect to *Nmax* parameter for both Kriging methods: however , a value from *Nmax*=10÷20 is recommended when using a Kriging interpolation method.

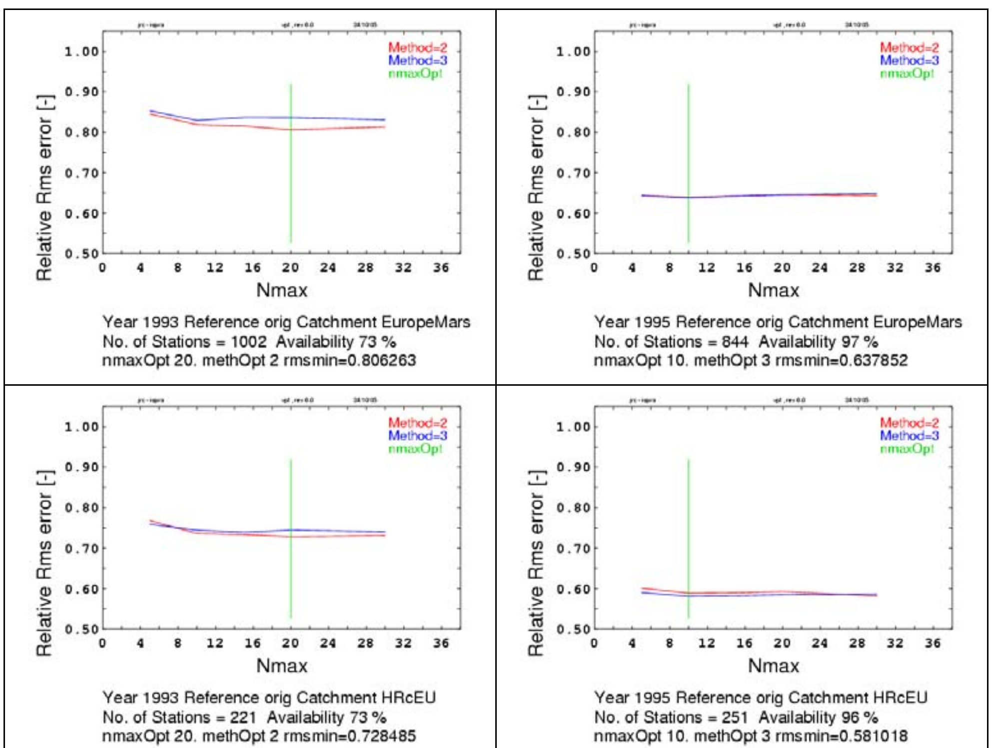


Fig 2.3.7 – Relative Difference calculated for the years 1993 and 1995: Method=2 stands for Ordinary Kriging while Method=3 stands for Cross Kriging

The parameter *Nmax* has then fixed to 15 and the calibration is continued in respect to *CleanFactor*= 0.75,1,1.5,2,3,4,5,10 (see section 2.3.1, step 1c for the *CleanFactor* meaning) . For *CleanFactor=10*, no cleaning is performed. No filtering is applied, i.e., the original data are used.

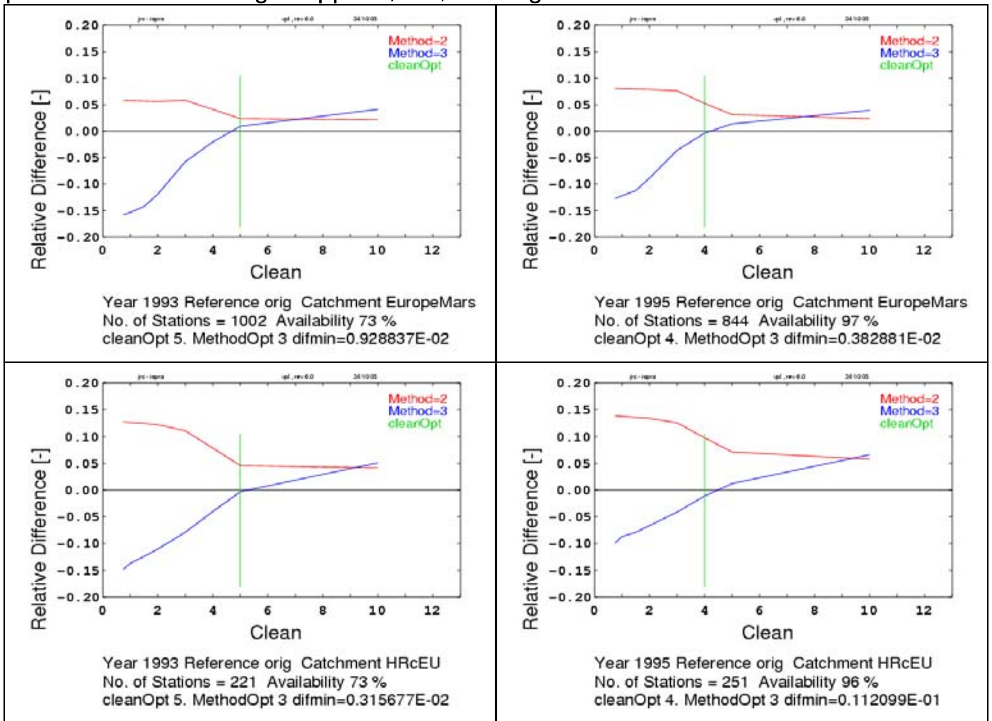


Fig 2.3.8 – Relative Difference calculated for the years 1993 and 1995: Method=2 stands for Ordinary Kriging while Method=3 stands for Cross Kriging

The following trends are identified (Fig. 2.3.8)

- In a cross validation the Cross Kriging method estimates less of the Ordinary Kriging method.
- For the Cross Kriging method the quantity *RelDiff* decreases with *CleanFactor*, while for the Ordinary Kriging method such quantity reaches a minimum for *CleanFactor*=4÷5.

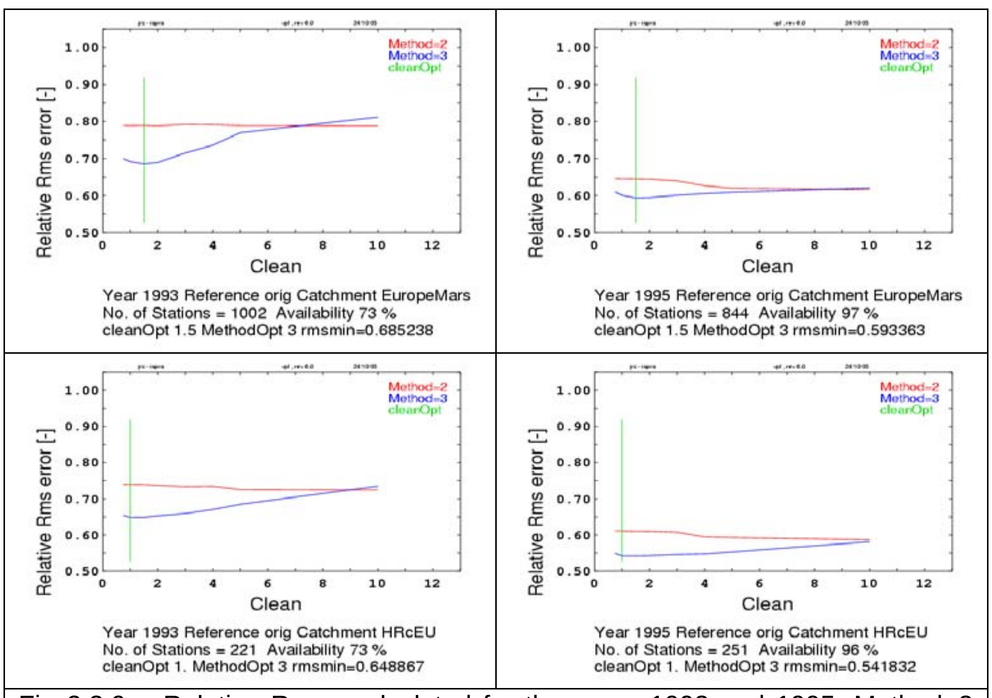


Fig 2.3.9 – Relative Rmse calculated for the years 1993 and 1995: Method=2 stands for Ordinary Kriging while Method=3 stands for Cross Kriging

The following trends are identified (Fig. 2.3.9)

- For “good data” like on year 2005 the effect of the *CleanFactor* is less strong as for “bad data” (year 1993).

For the Ordinary Kriging methods the quantity *RelRmse* has a minimum for *CleanFactor*=4 while for the Cross Kriging method continuously decreases with *CleanFactor.* but paid by an increase of the underestimation.

2.3.4 Cross Validation

Combining the trends observed during the calibration process, the following runs have been identified (see Tab. 3)

Mars Data Set	HR data set	*Idp*	*Nmax*	Method	*CleanFactor*
marsStatA	-	1	15	Inverse Distance	-
-	HrStatB	2	10	Inverse Distance	-
marsStatKord	HrStatKord	1	15	Ordinary Kriging	-
marsStatKordC4	HrStatKordC4	1	15	Ordinary Kriging	4
marsStatKcross	HrStatKcross	1	15	Cross Kriging	-
marsStatKcrossC4	HrStatKcrossC4	1	15	Cross Kriging	4

Tab. 3 – The selected parameters for the Kriging Methods validation

The cross validation have been performed for the years 1990-2004 using the Mars Station data and the HR rainfall data in Central Europe, producing finally the graphs for each area under observation. No filtering is applied, i.e., the original data are used.

In Fig 2.3.10 is shown that with the Ordinary Kriging method - without cleaning - an improvement is always realised for the *RelDiff* quantity paid by an increase of the *RelRmse*, while introducing the cleaning feature the results are worst for both quantities. With the Cross Kriging Method a lower *RelRmse* and a decreases on *RelDiff* is realised in case of cleaning.

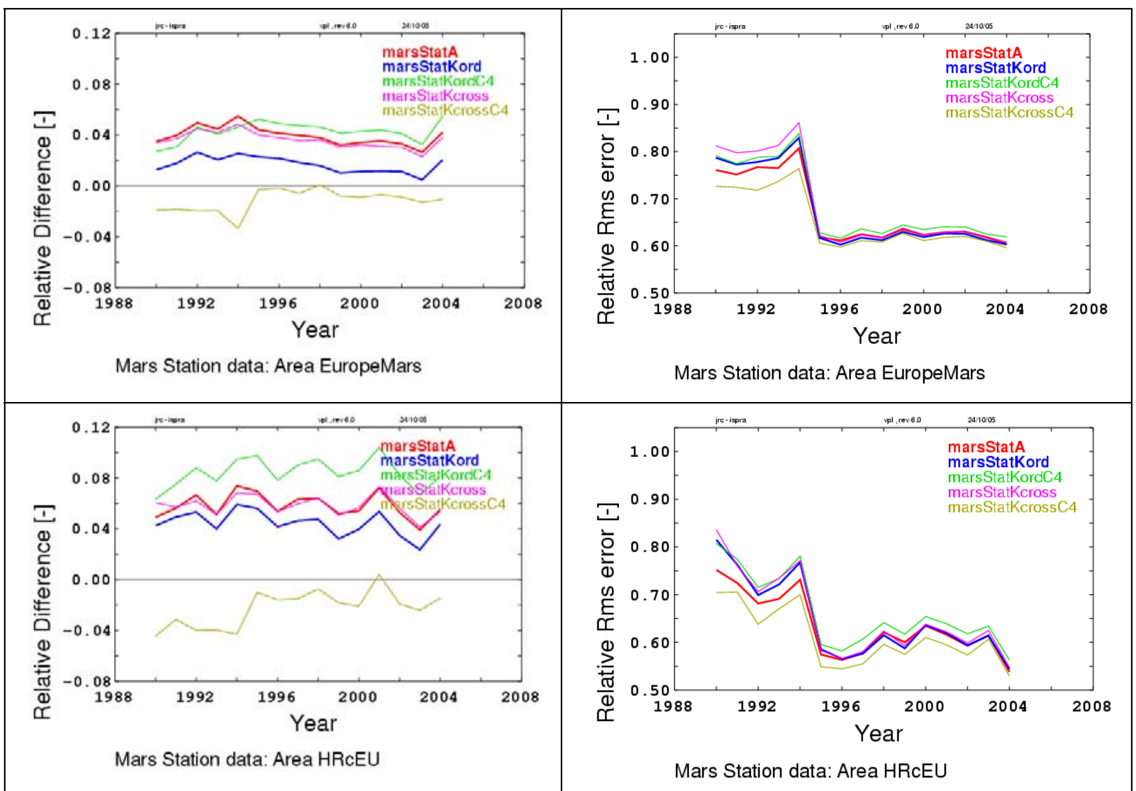


Fig 2.3.10 – Mars Station data plots: cross validation of the Kriging methods

Taking into account also the cross validation on the High resolution rainfall data (Fig. 2.3.11) the runs marsStatKord and marsStatKcrossC4 seem the more appropriate.

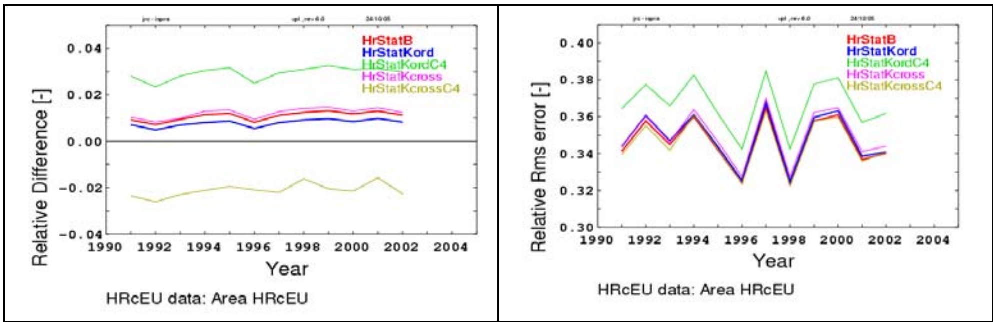


Fig 2.3.11 – High resolution rainfall data plots: cross validation of the Kriging method

2.3.5 Preliminary Maps Evaluation

In Fig.2.3.12 are shown, for the day 12 August 2002, the variogram models and the estimated standard error maps generated by the different Kriging methods and *CleanFactor* values. The standard error is higher in case of Ordinary Kriging, while it decreases for both methods by setting up a *CleanFactor*.

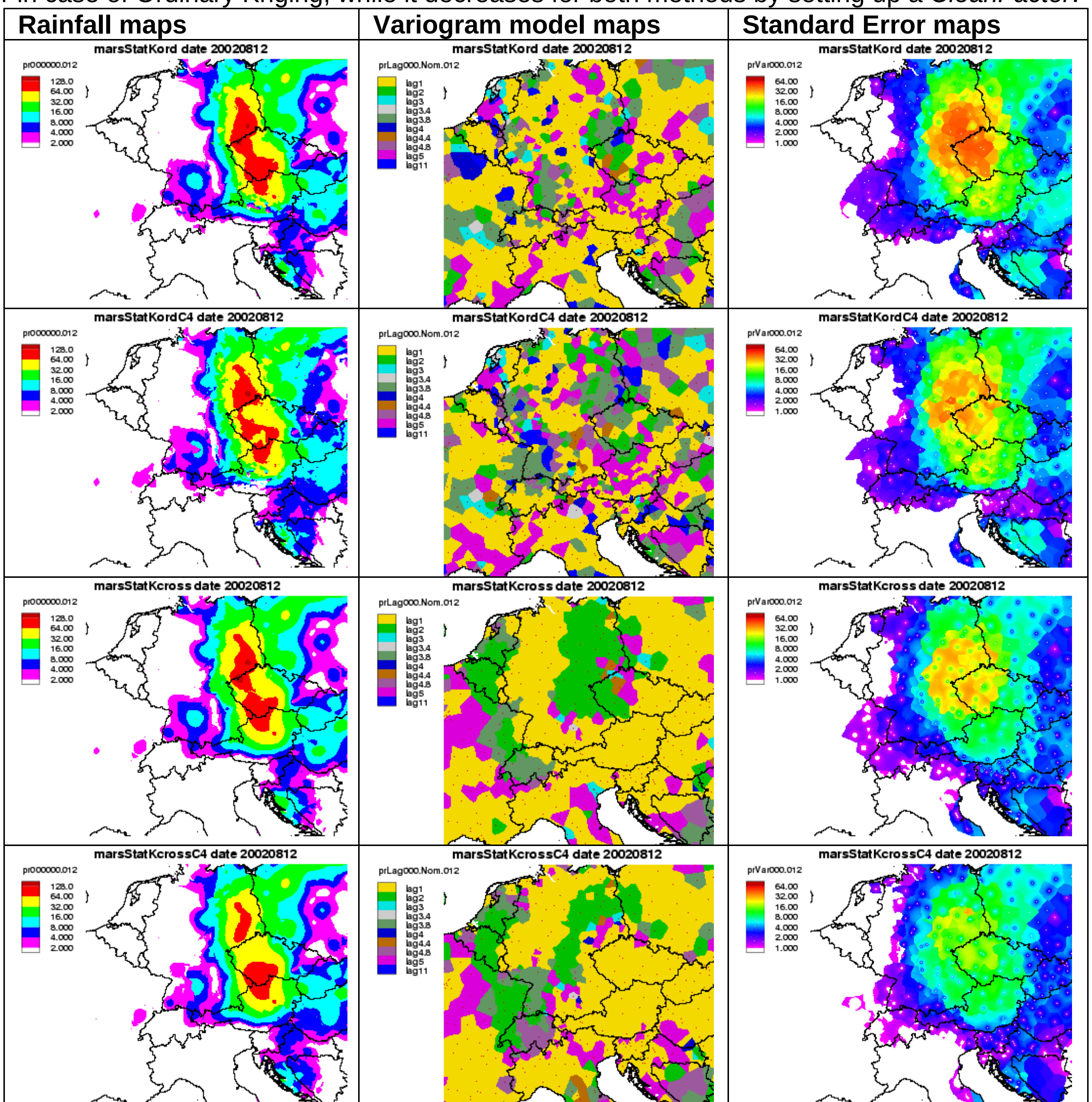


Fig. 2.3.12 – Rainfall, Variogram Models and Estimated Variance maps on 11 August 2002

For both methods the Linear Variogram Model is the most chosen. The variograms models with typlag= 3.0, 3.4, 4, 4.4 are selected only a few times (see Fig 2.3.13). Therefore, in order to save computational time are suggested the variograms models defined by *typlag* = 1, 2, 3.8, 5, 11

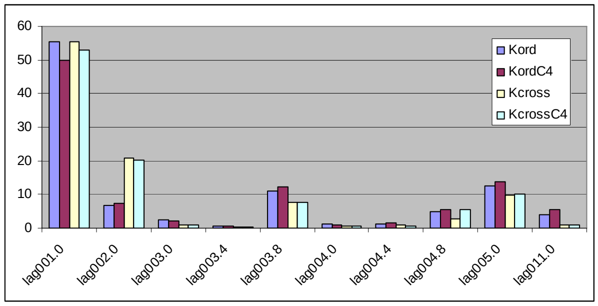


Fig. 2.3.13 – Variogram Model distribution

A methodology to validate the estimated standard error consists on comparing the error $\varepsilon_i = Z_i - Z_{rec,i}$ incurred during the cross validation with the estimated **standard error** σ_i and calculate [4]

$$x_i = \frac{\varepsilon_i}{\sigma_i} = \frac{Z_i - Z_{rec,i}}{\sigma_i}; \qquad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i; \qquad s_x = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)$$

Where *i* are the cells with station data and *N* are the no. of available stations. If the variogram models and the adopted kriging technique are appropriate, these mean $\bar{x}$ and **standard deviation** s_x should be approximately equal to 0 and 1 respectively. In Fig. 2.3.14 are shown the results of such comparisons.

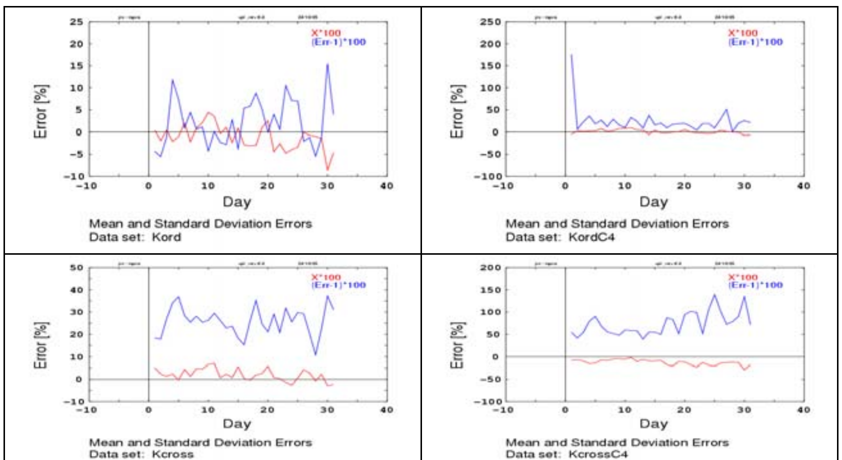


Fig 2.3.14 – Validation of the estimated standard error σ_i for the rainfall data on August 2002. In the legend X stand for $\bar{x}$ and Er stand for s_x

Only for the data sets *Kord* and *Kcross* (no cleaning) the mean error and the standard deviation from the error are acceptable, but applying the cleaning feature the evaluation of the standard error becomes unreliable. **The Ordinary Kriging method – without cleaning – is suggested when an evaluation on the standard error is requested**. In Fig 2.3.15 is shown that for low ε_i the standard error σ_i is overestimated for all Kriging methods.

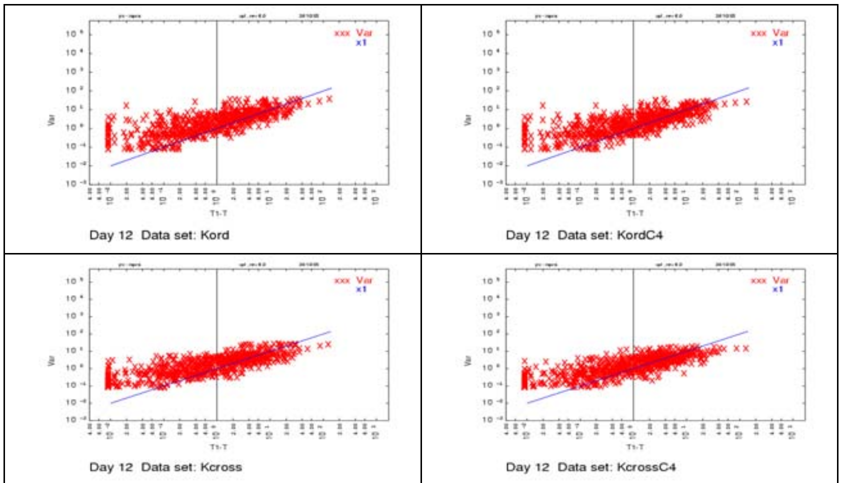


Fig 2.3.15 – Calculated error in the cross validation (x axis) vs estimated standard error (y axis) on 12 August 2002

2.3.6 Comparison of the Interpolated Data

The Mars and HR Station data are processed following the same procedure described in section 2.1.3. The mean annual rainfall is still underestimated (Fig 2.3.16) and maintains the same behavior for each data sets, but with a small improvement for the Ordinary Kriging data set and under estimation for Cross Kriging & Cleaning data set (marsStationKcrossC4). A small increase is seen in the years 1990-1994 for the *RelRmse* quantity (Fig 2.3.15). Both Kriging methods do not realize an improvement in respect to the Inverse Distance method: however, taking also into account the results on cross validation (section 2.3.5), the **Ordinary Kriging method – without cleaning - seems the more appropriate if the standard error estimation is requested.**

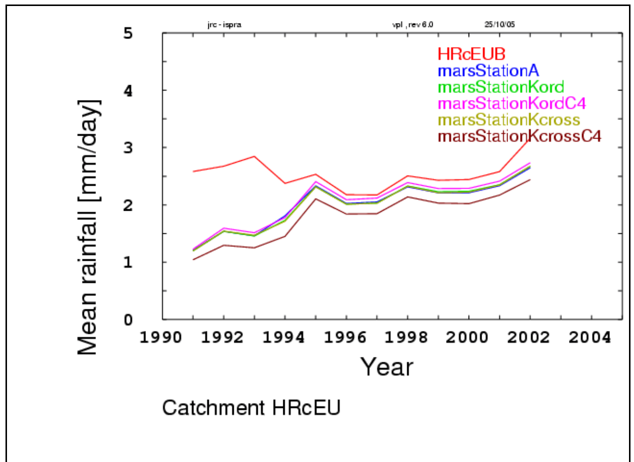


Fig 2.3.16– Mean annual rainfall: see Tab 3 for the legend explanation.

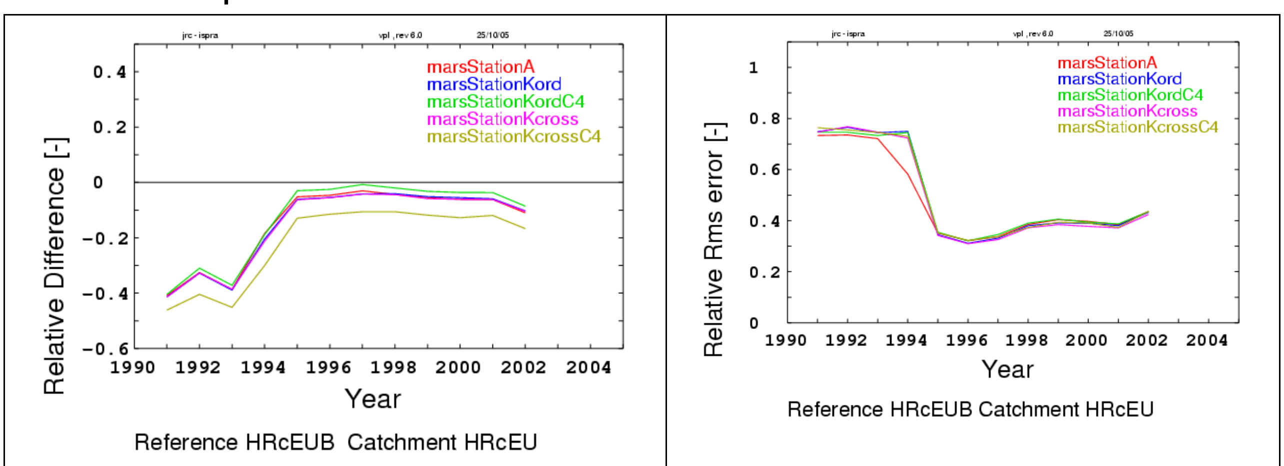


Fig 2.3.17 – Comparisons of mars Station data with the HR data set in Central Europe

In Fig. 2.3.18 are shown the rainfall maps related to the floods in Danube in August 2002. For such a period, **the data set generated by the Kriging methods have a pattern similar to the data set generated by the Inverse Distance Method, but with the capability to capture local high intensity rainfall**. Introducing the cleaning feature such patterns disappears.

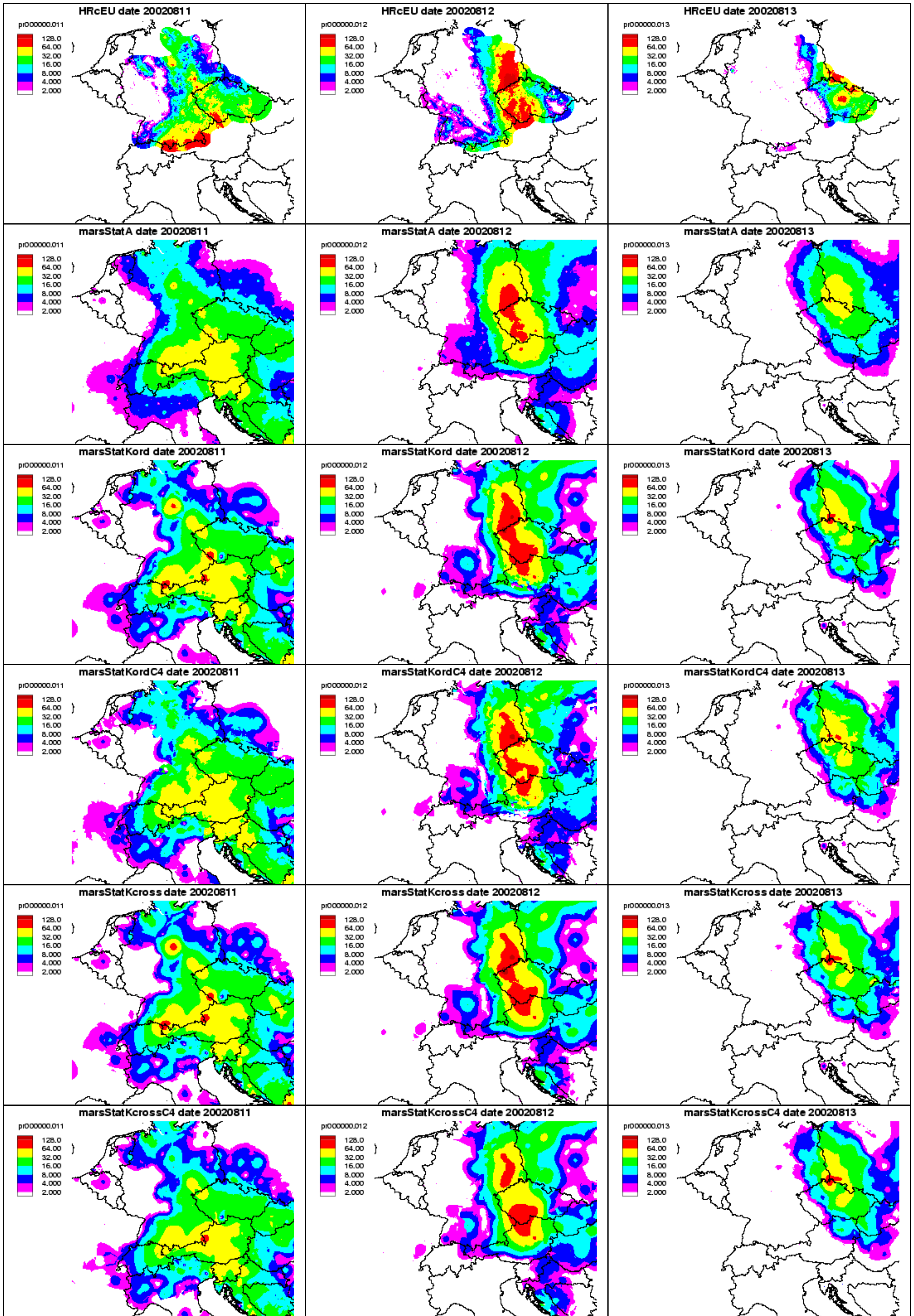


Fig. 2.3.18 – Rainfall maps for the period 11-13 August 2002

3. Filtering

In Section 1 was proposed a methodology to filter the suspicious values based on annual analysis of the rainfall intensity and data availability. This was done because some stations give always a very low value, or zero value. In this section is proposed a filtering process based on the cross validation results of the daily data sets. The advantage of such methodology is that an historical analysis is not necessary. The filtering process can be applied in all the interpolation methods above described.

After having recalculate all the station data, the Stations where the error is outside of a defined threshold are ignored according to the method of filtering chosen (see flag *typFilt*)

$$\begin{cases} Err_i > filterFactor & if \quad typFilt = 1 \\ abs(Err_i) > filterFactor & if \quad typFilt = 0 \\ Err_i < -filterFactor & if \quad typFilt = -1 \end{cases}$$

where

$$Err_i = \begin{cases} \dfrac{(Z_{i,rec} - Z_i)}{Rmse} & if \quad Rmse^2 < Z_i Z_{i,rec} \\[2ex] \dfrac{(Z_{i,rec} - Z_i)}{Z_i} & if \quad Rmse^2 > Z_i Z_{i,rec} \end{cases} \quad \text{and} \quad Rmse = \sqrt{\frac{\sum_{i=1,N} (Z_i - Z_{i,rec})^2}{N}}$$

The *Rmse* is evaluated in the area identified by *Radius*.

By the flag *typFilt* is possible to select between three methods of filtering:

- *typFilt*=1: only the Stations that are over estimated in a cross validation can be filtered
- *typFilt*=0: all the Stations can be filtered
- *typFilt*=-1: only the Stations that are under estimated in a cross validation can be filtered

The possibility of filtering only the Stations that are over estimated in a cross validation is introduced because - for any type of interpolation method analysed - the amount of rainfall is under estimated when compared with the high resolution data sets.

3.1 Calibration of the governing parameters

The calibration has done for the years 1993 (poor data) and 1995 (good data). Each station data is recalculated (source field) and compared with the original data (reference field) following the procedure described in Appendix 1.

The calibration is done in respect to the parameter *FilterFactor*= 0.25, 0.50, 0.75, 1, 1.5, 2, 3, 5. The filter is applied to the Inverse Distance Method, using 4 combinations of *Nmax*,*Idp* and *typFilt* parameters (see Tab. 4)

Data Set	*Idp*	*Nmax*	*typFilt*
filtA0	1	15	0
filtB0	2	10	0
filtA1	1	15	1
filtB1	2	10	1

Tab. 4 – The data sets used for the *filterFactor* calibration

For ***typFilt*=0** are identified the following trends (see Fig 3.1 and 3.2, data sets filtA0 and filtB0)

- the *RelDiff* quantity decreases for 1.5 < *filterFactor* < 10, reaches a maxima for *filterFactor* = 1 and decreases again
- the *RelRmse* decrease with *filterFactor*, reaching a minima for *filterFactor*<0.5

For ***typFilt*=1** are identified the following trends (see Fig 3.1 and 3.2, data sets filtA1 and filtB1)

- the *RelDiff* quantity starts to increase for *filterFactor* < 2
- the *RelRmse* quantity increases too. Values of *filterFactor* > 1 are recommended.

In both cases the effect of filtering is less important for “good data” (i.e. year1995).

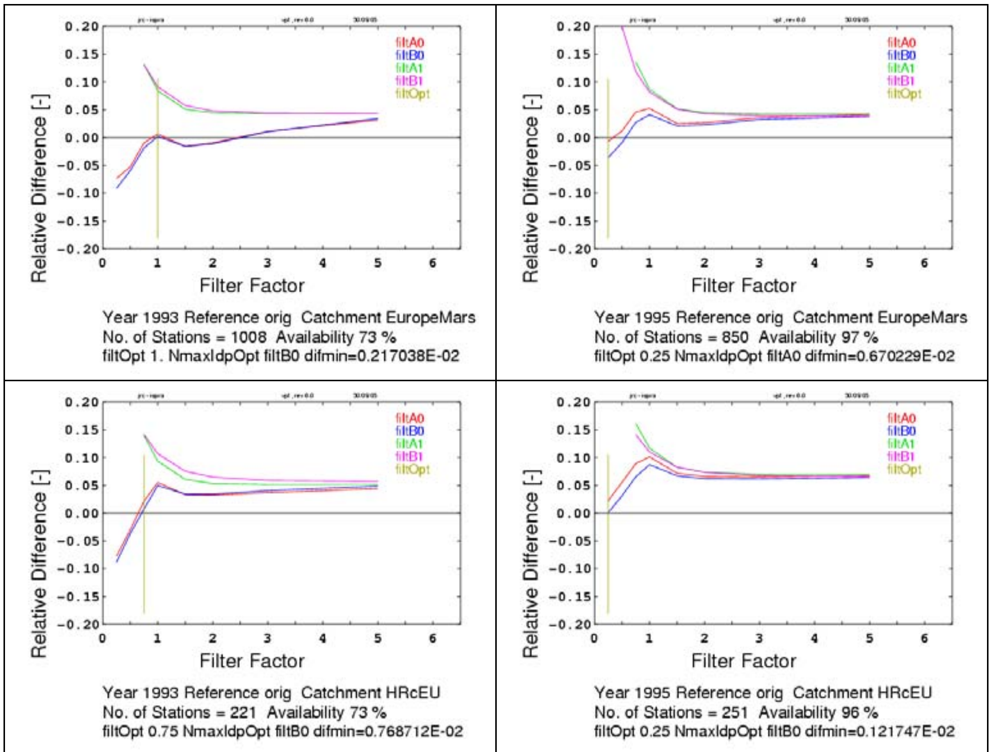


Fig 3.1 – Relative Difference calculated for the years 1993 and 1995

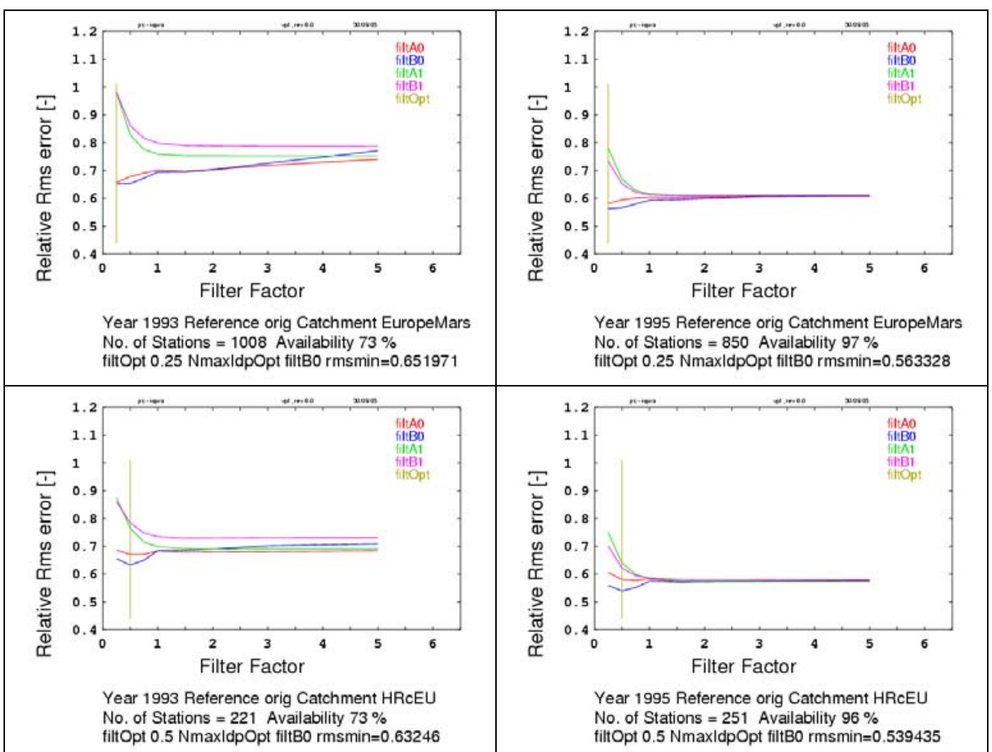


Fig 3.2 – Relative RMSE calculated for the years 1993 and 1995

3.2 Cross Validation

Combining the trends observed during the calibration process, the following runs have been identified (see Tab. 3)

Mars Data Set	HR data set	*Idp*	*Nmax*	Method	*FilterFactor*	*typFilt*
marsStatA	HrStatA	1	15	Inverse Distance	-	-
marsStatB	HrStatB	2	10	Inverse Distance	-	-
marsStatFilt2A0	HrStatFilt2A0	1	15	Inverse Distance	2	0
marsStatFilt2B0	HrStatFilt2B0	2	10	Inverse Distance	2	0
marsStatFilt1A0	HrStatFilt1A0	1	15	Inverse Distance	1	0
marsStatFilt1B0	HrStatFilt1B0	2	10	Inverse Distance	1	0

Tab. 5 – The selected parameters for the Filtering validation

With *filterFactor*=2 is realised an improvement in the *RelRmse* and *RelDiff* quantities , while with *filterFactor*=1 the *RelDiff* curves are shifted in the positive direction of 1 ÷ 2%. For the HR data sets the effect of filtering is negligible for *filterFactor*=2, while an overestimation of 1 ÷ 2% is also obtained for *filterFactor*=1. The data sets marsStatFilt2A0 and HrStatB seems the more appropriate.

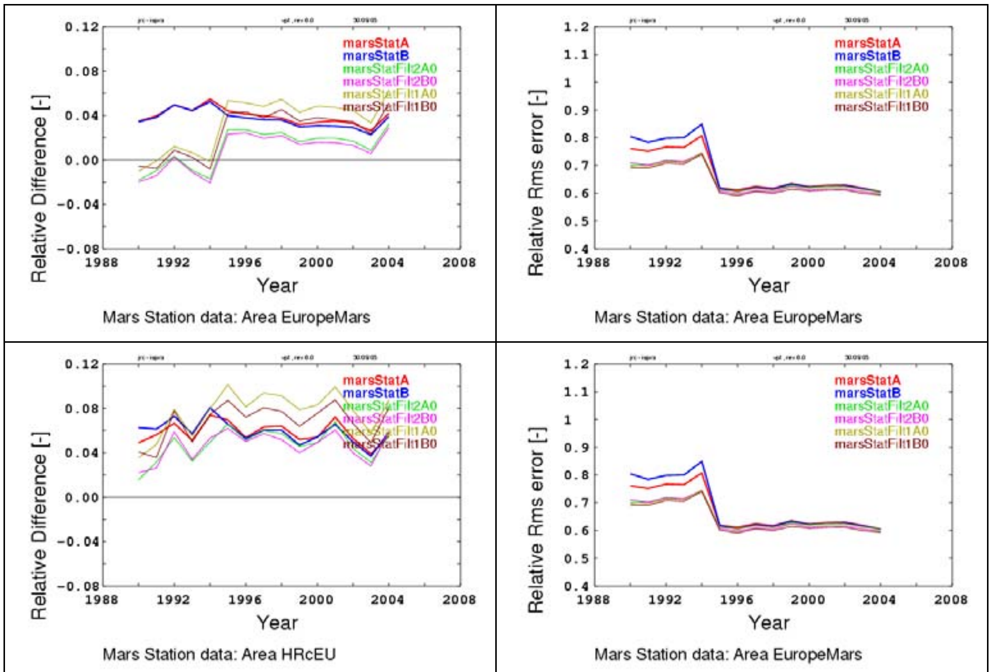


Fig. 3.3 – Mars Station data plots: filtering cross validation

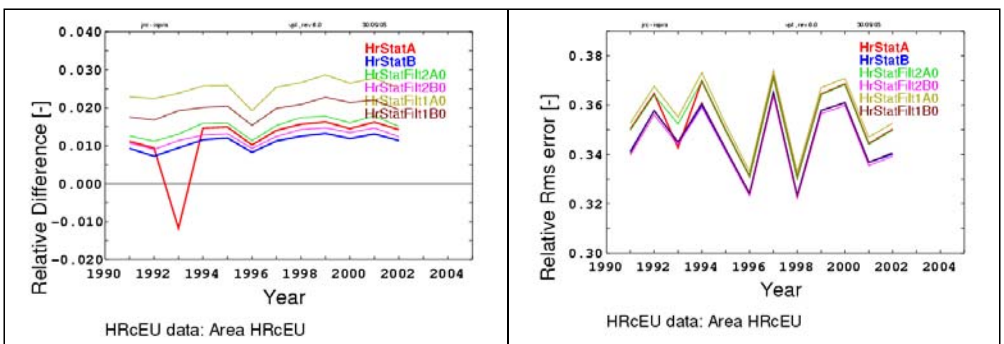


Fig. 3.4 – High Resolution rainfall data plots: filtering cross validation

3.3 Comparison of the Interpolated Data

The Mars and HR Station data are processed following the same procedure described in section 2.1.3. The mean annual rainfall is still underestimated (Fig 3.5 and 3.6) for the data sets with *filterFactor*=2, while a small improvement is obtained for *filterFactor*=1 mainly in the years 1995-2004. The data set marsStationFilt1A0 seems the more appropriate if one wants to compensate the underestimation of the Mars Station data in respect to the HR Station data, but taking into account that in a cross validation the more appropriate is the run marsStationFilt2A0.

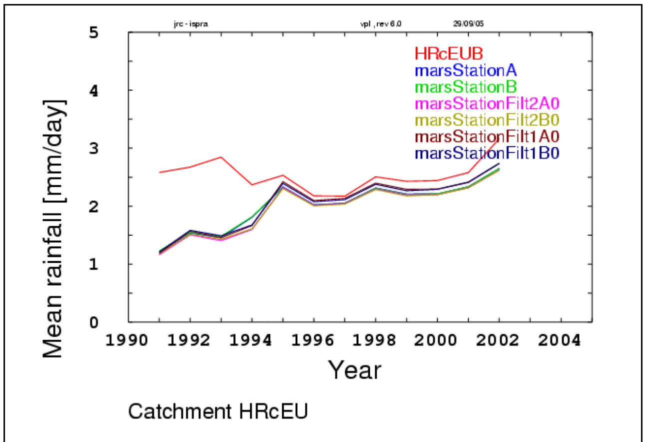


Fig 3.5 – Mean annual rainfall: see Tab 5 for the legend explanation."

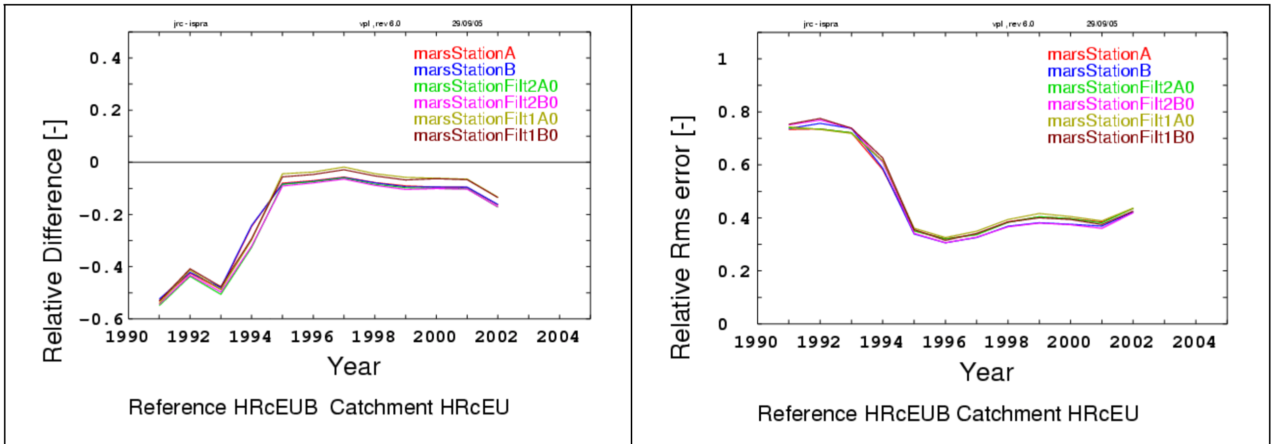


Fig 3.6 – Comparisons of mars Station data with the HR data set in Central Europe

In Fig 3.7 are shown the rainfall maps related to the floods in Danube on 11 August 2002: some isolated "spots" are removed.

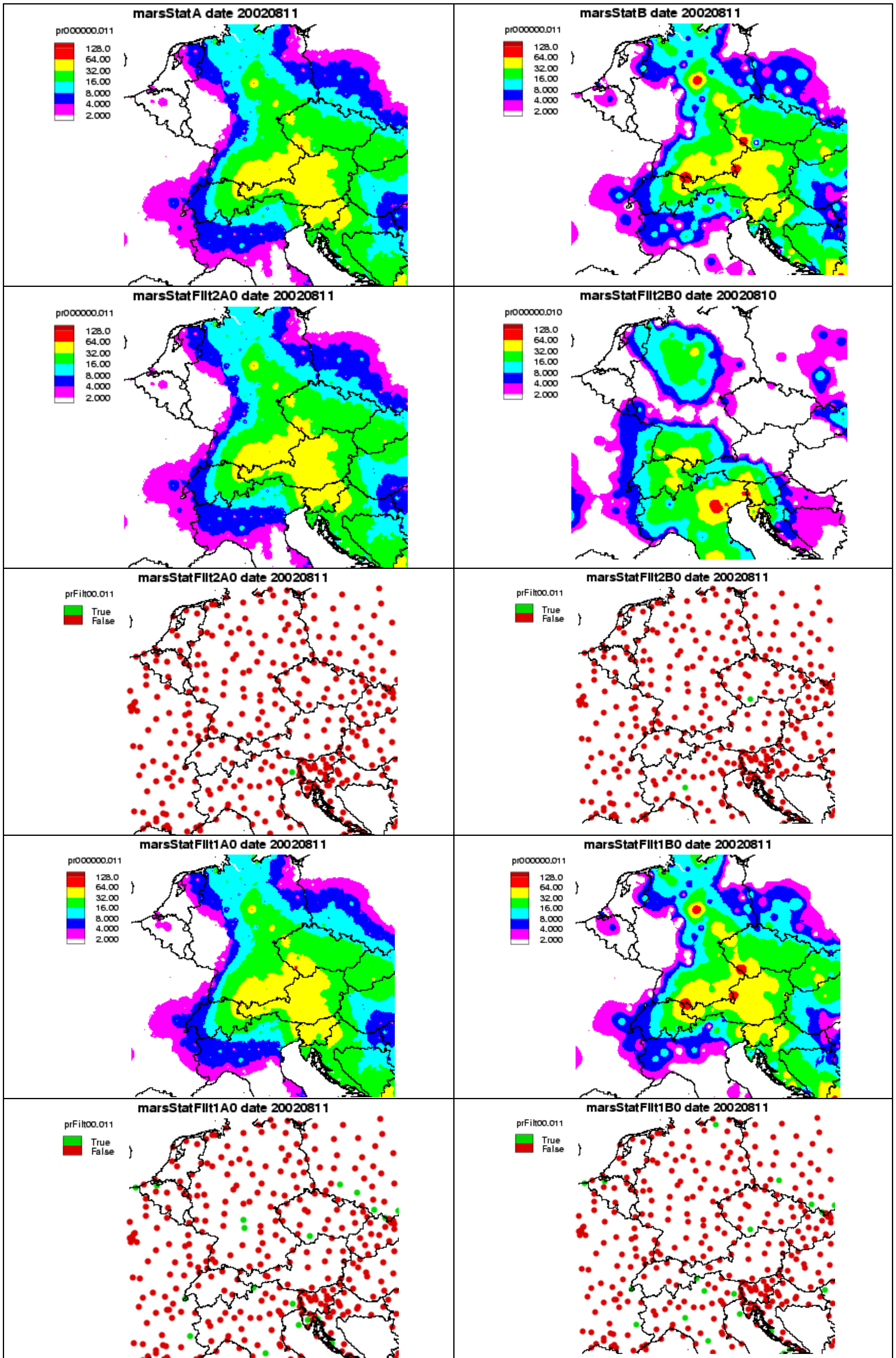


Fig 3.7 – Rainfall and filter maps on 11 August 2002. The filtering process eliminates isolated spots. The effect of filtering is more evident with lower *Nmax* parameter.

In Fig 3.8 is shown that only few stations are filtered in case of *filterFactor*=2, while the percentage increases in case of *filterFactor*=1. For the Mars Station, the percentage is higher for the years 1990-1994 if *filterFactor*=2 is chosen. This indicates that with such values the filtering process is able to correct in a certain extend the "bad" data of years 1990-1994. For the High resolution data sets the percentage of filtered stations is about 50 % of the Mars Stations. The no. of filtered stations increases if *Nmax* decreases. The *filterFactor=2* is suggested when a daily filter is applied.

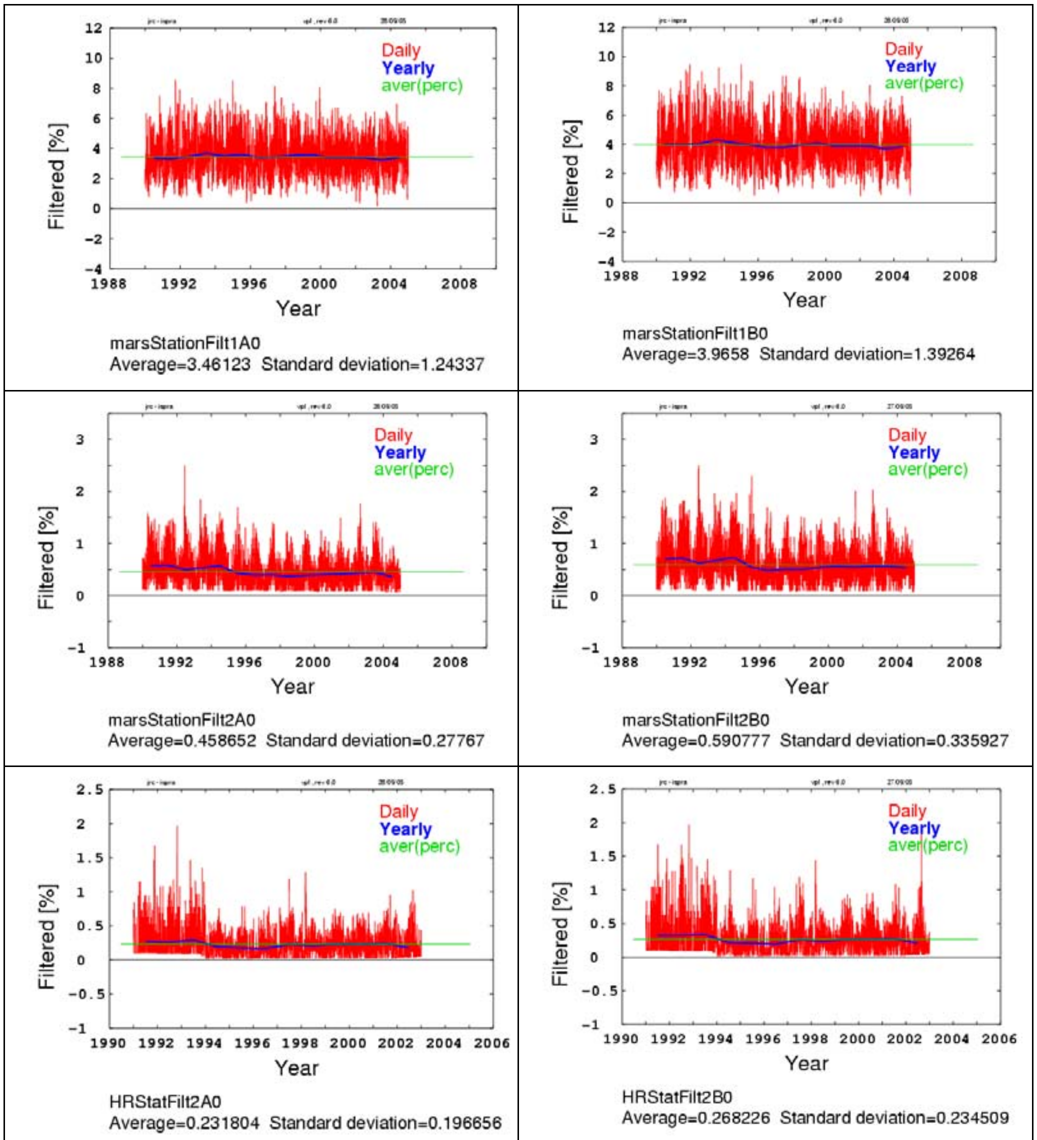


Fig. 3.8 – Filtered stations

4. Weighting

In this section is proposed a methodology on weighting (and filtering) the suspicious values based on the analysis of the annual rainfall and data availability. The proposed methodology differs from that proposed in section 1 by the following features

- the minimum and maximum threshold are not global values, but are space defined maps evaluated by a statistical analysis of the available data
- the weight is proportional to the no. of days availability in the year

The procedure adopted to generate the weight maps is the following:

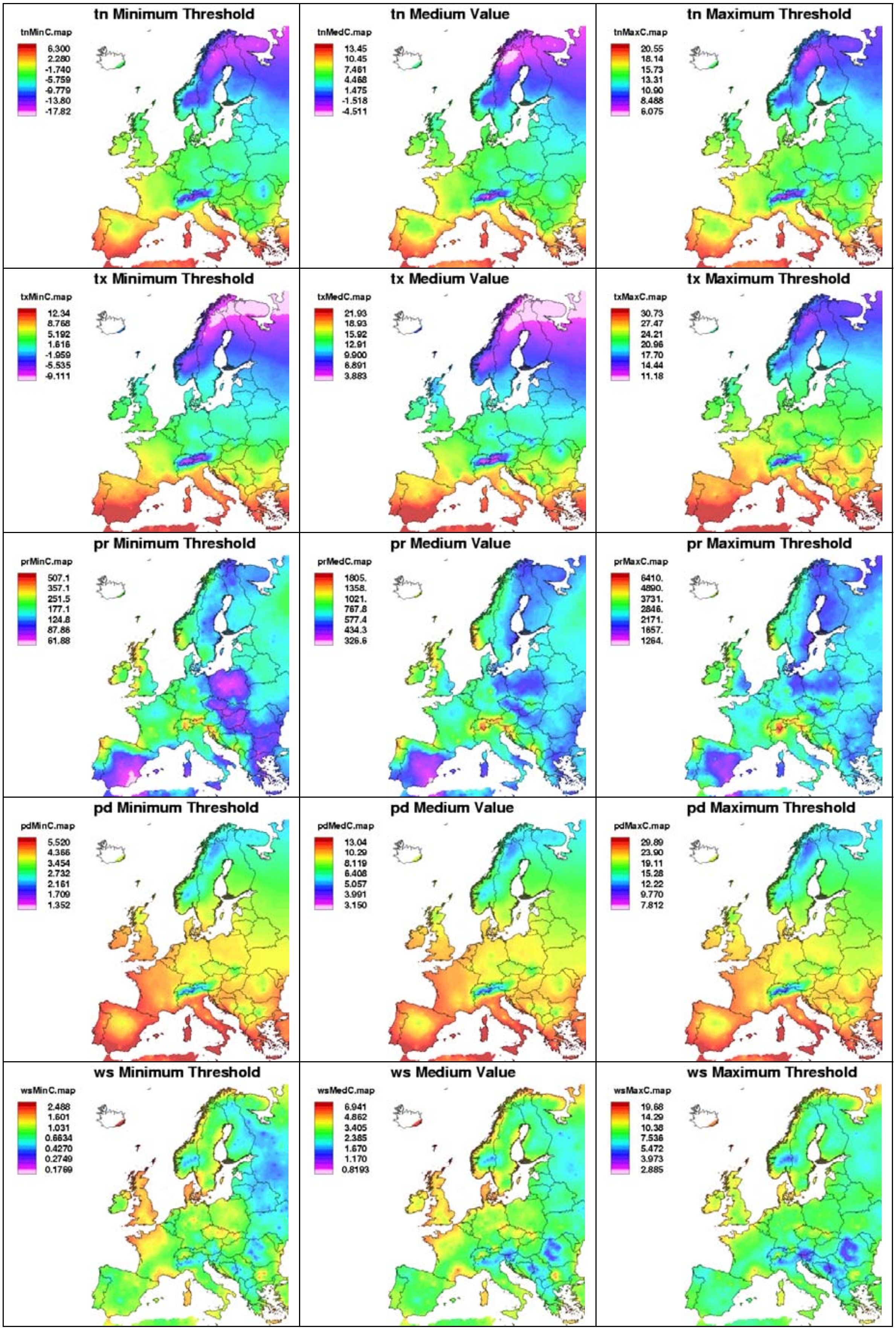
1) Maps on minimum and maximum annual mean thresholds are evaluated taking into account the annual data in the years 1990-2004 (see Fig 4.1)
    a. Only the Stations with annual availability higher of 80% are considered
    b. The maps on minimum and maximum annual values over the 15 years are updated in order to obtain thresholds with a range higher of what obtained in the period under analysis, i.e., the minimum value maps are divided and the maximum value maps are multiplied respectively by a factor that depends on the ratio between minimum and maximum value
    c. Such minima and maxima are "smoothed", following the procedure described in Section 3: *filterFactor*=2 and *typFilt*=0 are adopted.
2) Annual weight maps are set proportional to the no. of days availability, but set to zero if
    a. the no. of days availability is lower than 10
    b. the annual mean values are outside of the minimum and maximum threshold values as calculated at step 1
    c. the cross validation of the annual mean lead to filter the station as described in Section 3

Note that the weights are used only in case of inverse distance and fitting methods. By imposing a zero value correspond to exclude the station: in this case the filtering is applied to all methods. The values for *filterFactor, typFilt* and minimum of days availability have been chosen in a way that the resulting suspicious values is not higher of few percentage and in the same time an improvement in respect to the high resolution data is realised.

The same procedure is applied to the remaining variables available in the mars Station data base (see Tab. 6), obtaining the maps in Fig 4.1. The variables *tn, tx, pd, e0, es, et* are "corrected" in respect to the difference in altitude from the Station and the cell center [3]. Therefore, the cross validation of the interpolation methods applied to the above mentioned meteo variables includes also the cross validation of such correction. Note that the variables *pd, ws* are used only in case the correction of the potential evaporation quantities (*e0,es,et*) is requested. The minimum and maximum maps values must be considered as thresholds and not the minimum and maximum values observed in the years 1990-2004.

Variable	Value	Description
tn	MINIMUM_TEMPERATURE	minimum temperature (°C)
tx	MAXIMUM_TEMPERATURE	maximum temperature (°C)
pr	RAINFALL	mean daily rainfall (mm)
pd	VAPOUR_PRESSURE	mean daily vapour pressure (hPa)
ws	WINDSPEED	mean daily windspeed at 10m (m/s)
e0	E0	Penman potential evaporation from a free water surface (mm/day)
es	ES0	Penman potential evaporation from a moist bare soil surface (mm/day)
et	ET0	Penman potential transpiration from a crop canopy (mm/day)
cr	CALCULATED_RADIATION	daily global radiation in $KJ/m^2/day$

Tab. 6 – The Mars Station Variables


tn Minimum Threshold
tnMinC.map
6.300
2.280
-1.740
-5.759
-9.779
-13.80
-17.82
tn Medium Value
tnMedC.map
13.45
10.45
7.461
4.468
1.475
-1.518
-4.511
tn Maximum Threshold
tnMaxC.map
20.55
18.14
15.73
13.31
10.90
8.488
6.075
tx Minimum Threshold
txMinC.map
12.34
8.768
5.192
1.616
-1.959
-5.535
-9.111
tx Medium Value
txMedC.map
21.93
18.93
15.92
12.91
9.900
6.891
3.883
tx Maximum Threshold
txMaxC.map
30.73
27.47
24.21
20.96
17.70
14.44
11.18
pr Minimum Threshold
prMinC.map
507.1
357.1
251.5
177.1
124.8
87.88
61.88
pr Medium Value
prMedC.map
1805.
1358.
1021.
767.8
577.4
434.3
326.6
pr Maximum Threshold
prMaxC.map
6410.
4890.
3731.
2846.
2171.
1657.
1264.
pd Minimum Threshold
pdMinC.map
5.520
4.366
3.454
2.732
2.161
1.709
1.352
pd Medium Value
pdMedC.map
13.04
10.29
8.119
6.408
5.057
3.991
3.150
pd Maximum Threshold
pdMaxC.map
29.89
23.90
19.11
15.28
12.22
9.770
7.812
ws Minimum Threshold
wsMinC.map
2.488
1.601
1.031
0.6634
0.4270
0.2749
0.1769
ws Medium Value
wsMedC.map
6.941
4.862
3.405
2.385
1.670
1.170
0.8193
ws Maximum Threshold
wsMaxC.map
19.68
14.29
10.38
7.536
5.472
3.973
2.885

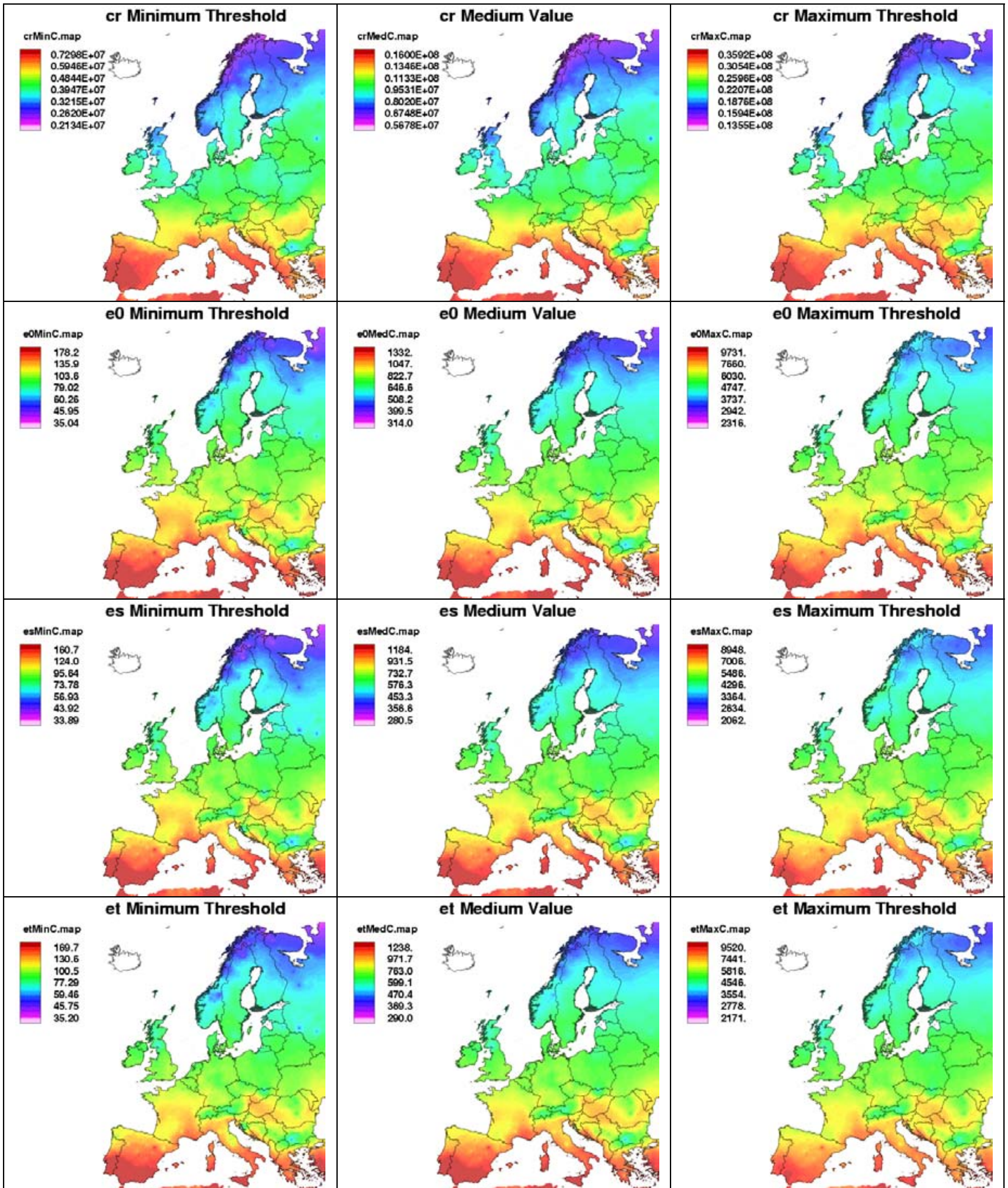
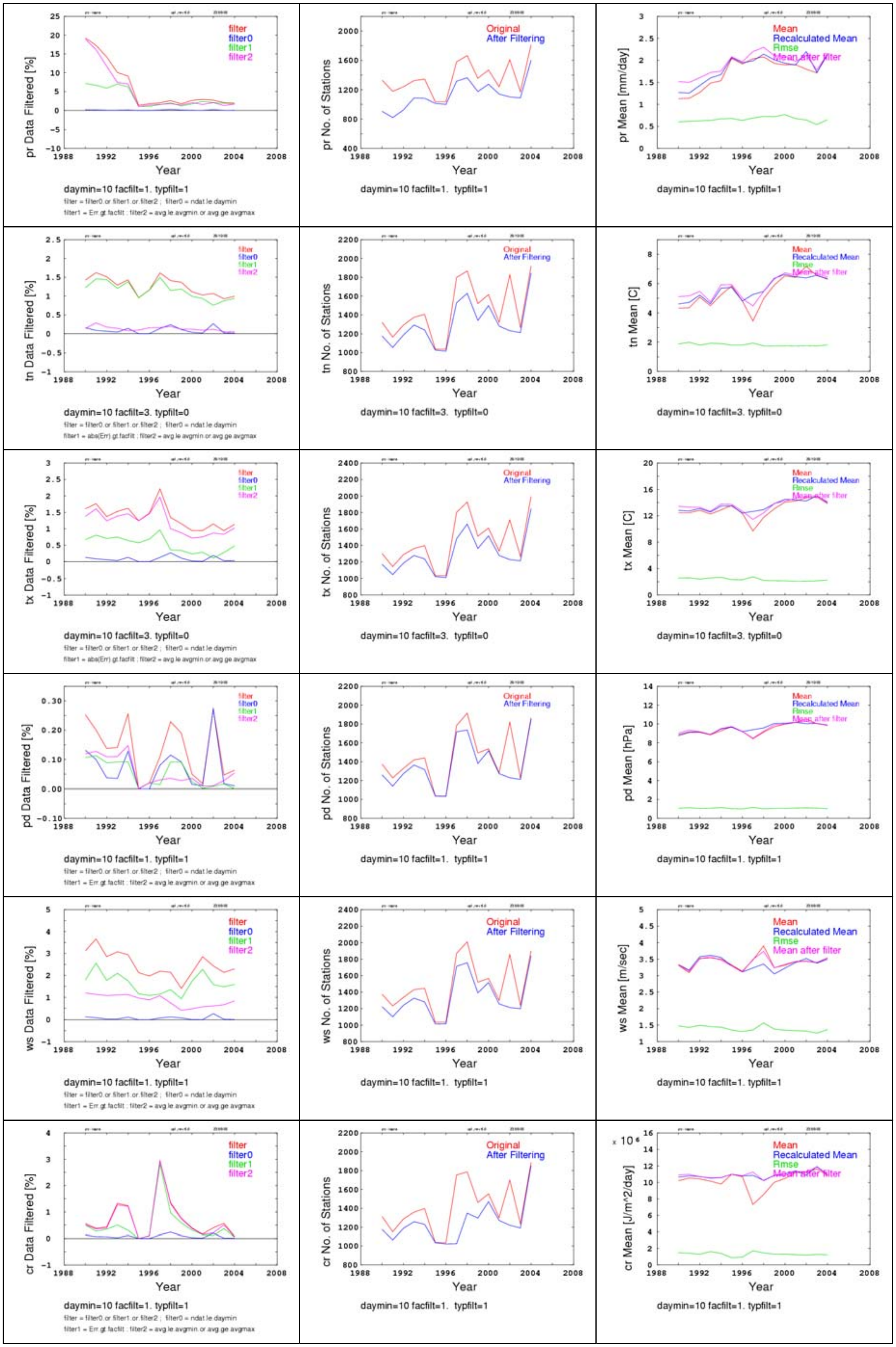


Fig. 4.1 – Minimum threshold, medium value and maximum threshold obtained after 1990-2004 analysis

The results of applying the procedure to generate the yearly weight maps -as described in step 2- are summarised in Fig. 4.2. The no. of filtered station is in the order of 1-3% for all the variables apart of the rainfall data set, where for the years 1990-1994 is from 20% to 10%. In the years 1997-1998 some stations are filtered because the no. of days availability is below the threshold of 10 days in the year, causing a filtering of less then 0.5 % of the data. After filtering the mean values are more regular.


pr Data Filtered [%]
Year
filter
filter0
filter1
filter2
daymin=10 facfilt=1. typfilt=1
filter = filter0 .or filter1 .or filter2 ; filter0 = ndat.le.daymin
filter1 = Err.gt.facfilt ; filter2 = avg.le.avgmin.or.avg.ge.avgmax
pr No. of Stations
Original
After Filtering
daymin=10 facfilt=1. typfilt=1
pr Mean [mm/day]
Mean
Recalculated Mean
Rmse
Mean after filter
daymin=10 facfilt=1. typfilt=1
tn Data Filtered [%]
daymin=10 facfilt=3. typfilt=0
filter = filter0 .or filter1 .or filter2 ; filter0 = ndat.le.daymin
filter1 = abs(Err).gt.facfilt ; filter2 = avg.le.avgmin.or.avg.ge.avgmax
tn No. of Stations
daymin=10 facfilt=3. typfilt=0
tn Mean [C]
daymin=10 facfilt=3. typfilt=0
tx Data Filtered [%]
daymin=10 facfilt=3. typfilt=0
filter = filter0 .or filter1 .or filter2 ; filter0 = ndat.le.daymin
filter1 = abs(Err).gt.facfilt ; filter2 = avg.le.avgmin.or.avg.ge.avgmax
tx No. of Stations
daymin=10 facfilt=3. typfilt=0
tx Mean [C]
daymin=10 facfilt=3. typfilt=0
pd Data Filtered [%]
daymin=10 facfilt=1. typfilt=1
filter = filter0 .or filter1 .or filter2 ; filter0 = ndat.le.daymin
filter1 = Err.gt.facfilt ; filter2 = avg.le.avgmin.or.avg.ge.avgmax
pd No. of Stations
daymin=10 facfilt=1. typfilt=1
pd Mean [hPa]
daymin=10 facfilt=1. typfilt=1
ws Data Filtered [%]
daymin=10 facfilt=1. typfilt=1
filter = filter0 .or filter1 .or filter2 ; filter0 = ndat.le.daymin
filter1 = Err.gt.facfilt ; filter2 = avg.le.avgmin.or.avg.ge.avgmax
ws No. of Stations
daymin=10 facfilt=1. typfilt=1
ws Mean [m/sec]
daymin=10 facfilt=1. typfilt=1
cr Data Filtered [%]
daymin=10 facfilt=1. typfilt=1
filter = filter0 .or filter1 .or filter2 ; filter0 = ndat.le.daymin
filter1 = Err.gt.facfilt ; filter2 = avg.le.avgmin.or.avg.ge.avgmax
cr No. of Stations
daymin=10 facfilt=1. typfilt=1
x 10^6
cr Mean [J/m^2/day]
daymin=10 facfilt=1. typfilt=1

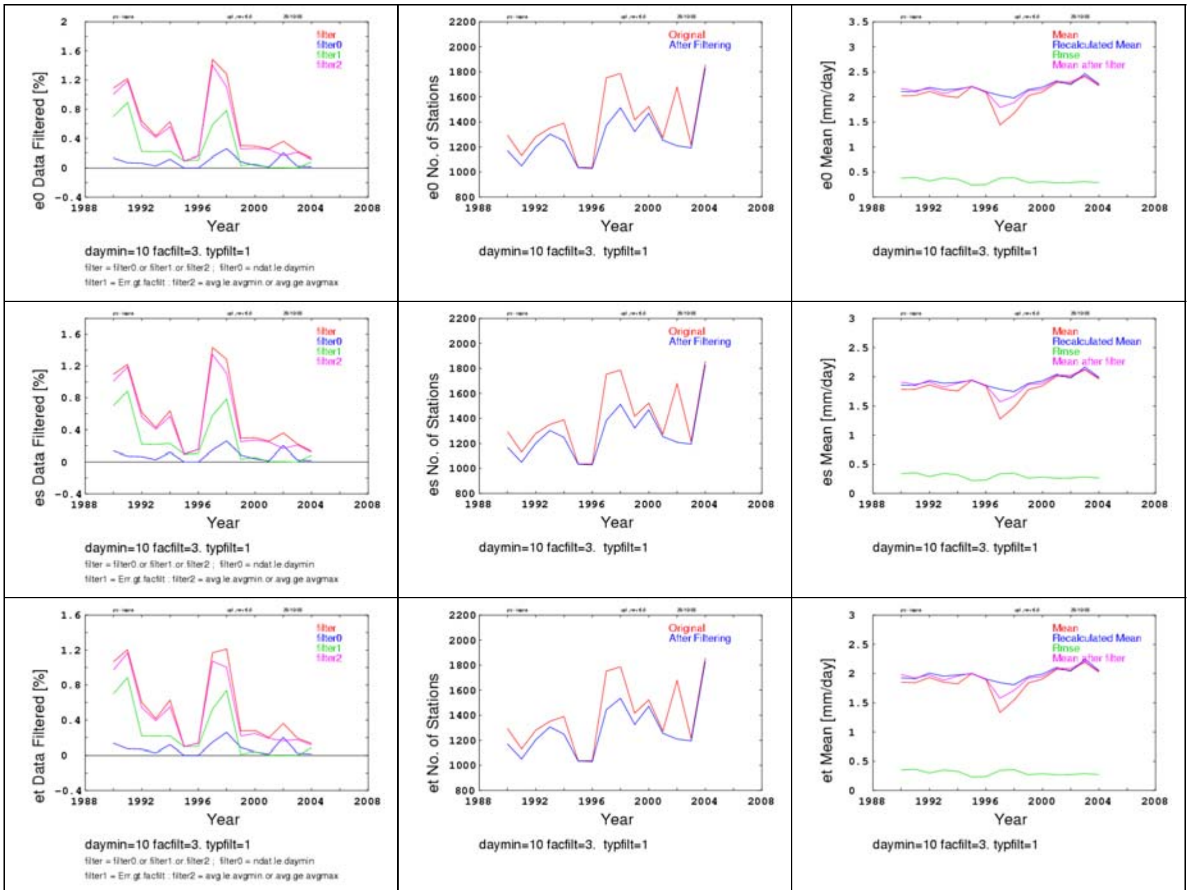


Fig. 4.2 – Pre-analysis on the effect of weighting for the Mars Station variables

4.1 Cross Validation

The following runs have been identified (see Tab. 7) for the cross Validation of Filtering & Weighting processes. Weight maps are also generated for the High Resolution data sets, following the same procedure as applied to the Mars Station data.

Mars Data Set	HR data set	*Idp*	*Nmax*	Method	*FilterFactor*	*typFilt*	Weight
marsStatA	-	1	15	Inverse Distance	-	-	No
marsStatAW	-	1	15	Inverse Distance	-	-	Yes
-	HrStatB	2	10	Inverse Distance	-	-	No
-	HrStatBW	2	10	Inverse Distance	-	-	Yes
marsStatWFilt2A0	HrStatWFilt2A0	1	15	Inverse Distance	2	0	Yes
marsStatWFilt2B0	HrStatWFilt2B0	2	10	Inverse Distance	2	0	Yes
marsStatWFilt1A0	HrStatWFilt1A0	1	15	Inverse Distance	1	0	Yes
marsStatWFilt1B0	HrStatWFilt1B0	2	10	Inverse Distance	1	0	Yes

Tab. 7 – The selected parameters for the Filtering & Weighting validation

Including the weight maps an improvement is seen for the *RelRmse* as well as the *RelDiff* quantities (see Fig. 4.3 and 4.4). A further improvement in the *RelRmse* quantity is realised if the filtering process is included too, while the effect of combining filtering & weighting is less strong in the *RelDiff* quantity (see also Fig. 3.3, where only filtering is applied). The data sets marsStatAW (only weighting), marsStatWFilt2A0 (filtering & weighting) and HrStatBW seems the more appropriate.

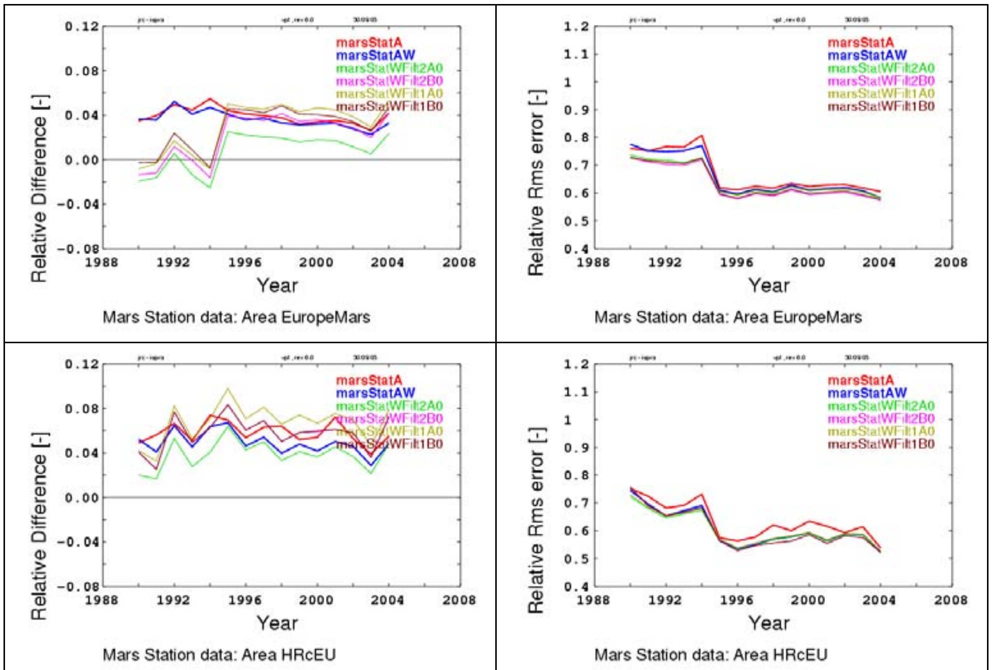


Fig. 4.3 – Mars Station data plots: filtering & weighting cross validation

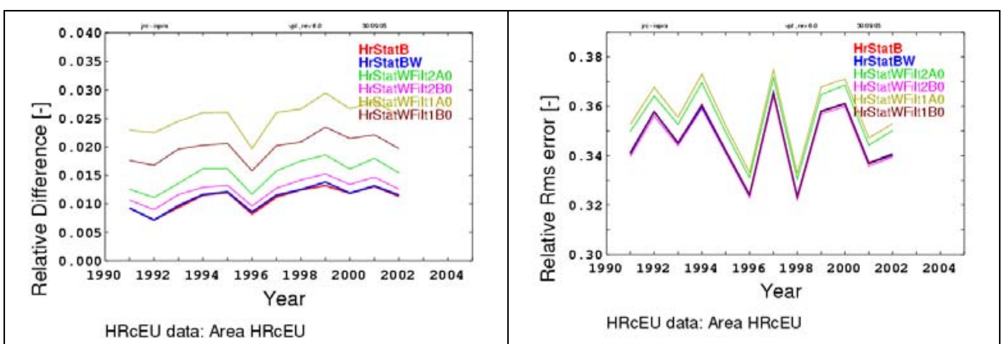


Fig. 4.4 – High Resolution rainfall data plots: filtering & weighting cross validation

4.2 Data Comparison

The Mars and HR Station data are processed following the procedure described in section 2.1.3. By the weighting process a gain of +10% in the total rainfall estimation and a few percentage in the *RelRmse* is obtained for the years 1990-1994 (see Fig. 4.5 and 4.6). Combining filtering & weighting an improvement is also obtained for the years 1995-2004 in case *filterFactor*=1 is defined (data set marsStationWFilt1A0), but with an increase in the underestimation for the years 1990-1994 in a comparison with only weighting process (data set marsStationAW). In this comparison the run marsStationAW seems the more appropriate.

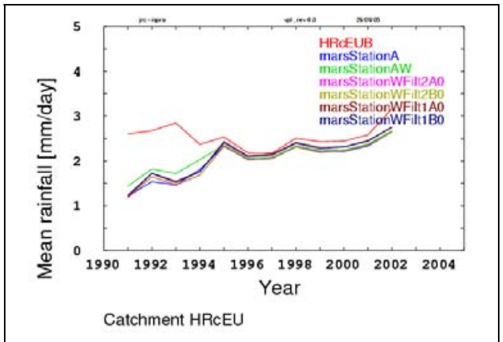


Fig 4.5 – Mean annual rainfall: see Tab 5 for the legend explanation.”

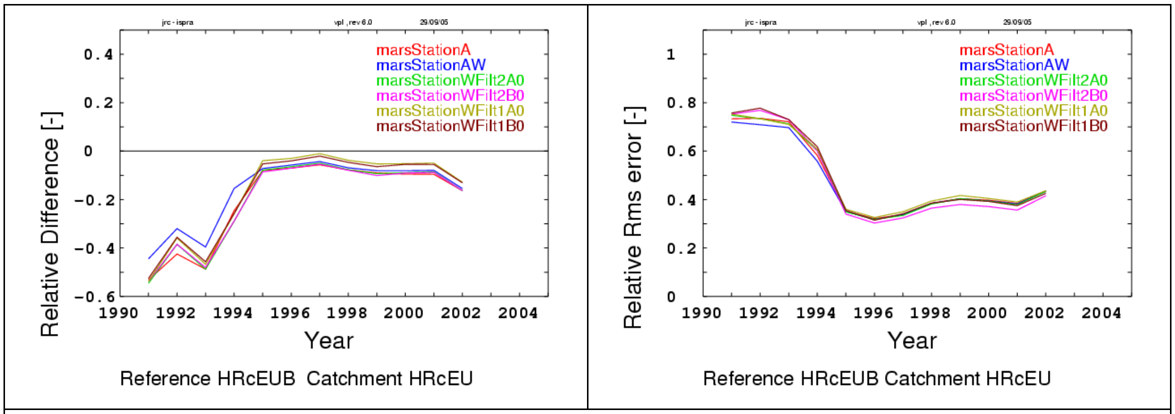


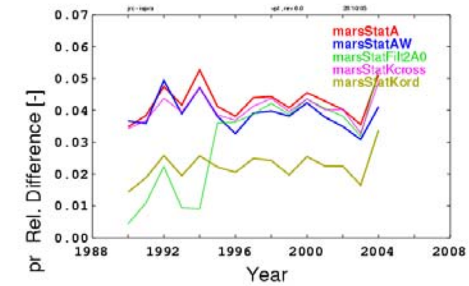
Fig 4.6 – Comparisons of mars Station data with the HR data set in Central Europe

5 - Cross Validation of the full set of meteo variables

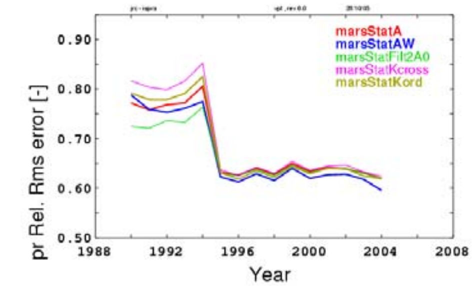
The cross validation of different interpolation methods combined with filtering & weighting is done for the meteo variables described in Tab. 6, applying the best combination parameters identified in the previous sections (see Tab. 8).

Mars Data Set	*Idp*	*Nmax*	Method	*FilterFactor*	*typFilt*	CleaningFactor	Weight
marsStatA	1	15	Inverse Distance	-	-	-	No
marsStatAW	1	15	Inverse Distance	-	-	-	Yes
marsStatFilt2A0	1	15	Inverse Distance	2	0	-	No
marsStatKord	-	15	Ordinary Kriging	-	-	-	No
marsStatKcross	-	15	Cross Kriging	-	-	-	No

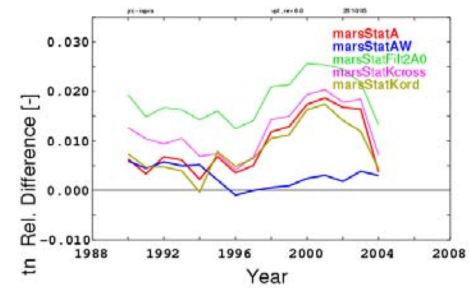
Tab. 8 – The selected parameters for the meteo variables cross validation



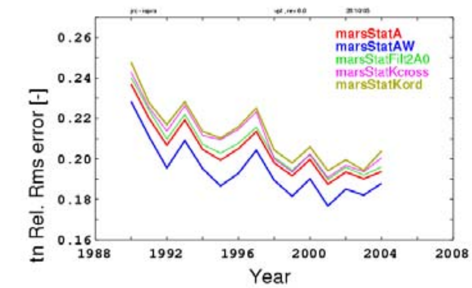
Mars Station Data: Area EuropeMars Variable pr



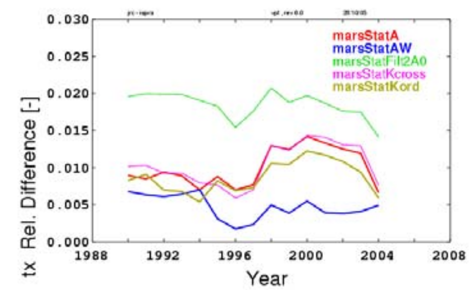
Mars Station Data: Area EuropeMars Variable pr



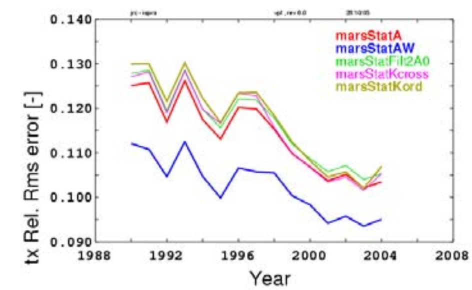
Mars Station Data: Area EuropeMars Variable tn



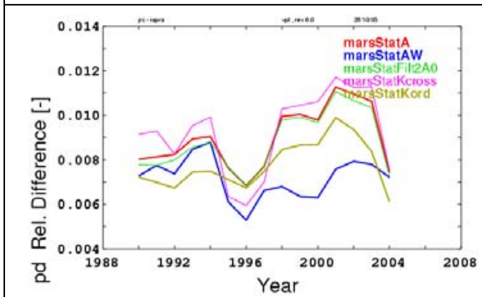
Mars Station Data: Area EuropeMars Variable tn



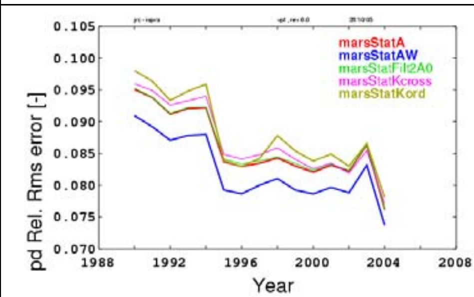
Mars Station Data: Area EuropeMars Variable tx



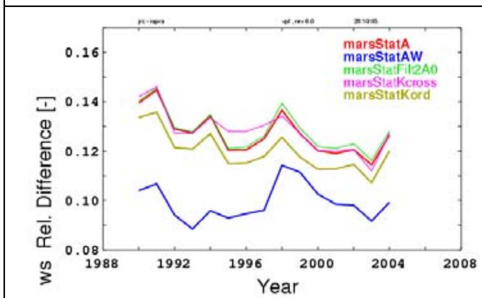
Mars Station Data: Area EuropeMars Variable tx



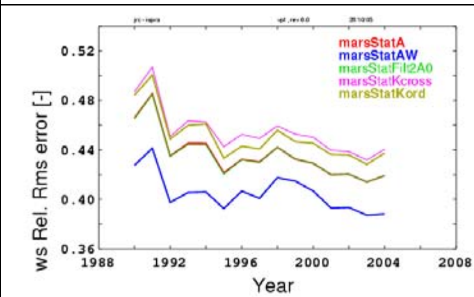
Mars Station Data: Area EuropeMars Variable pd



Mars Station Data: Area EuropeMars Variable pd



Mars Station Data: Area EuropeMars Variable ws



Mars Station Data: Area EuropeMars Variable ws

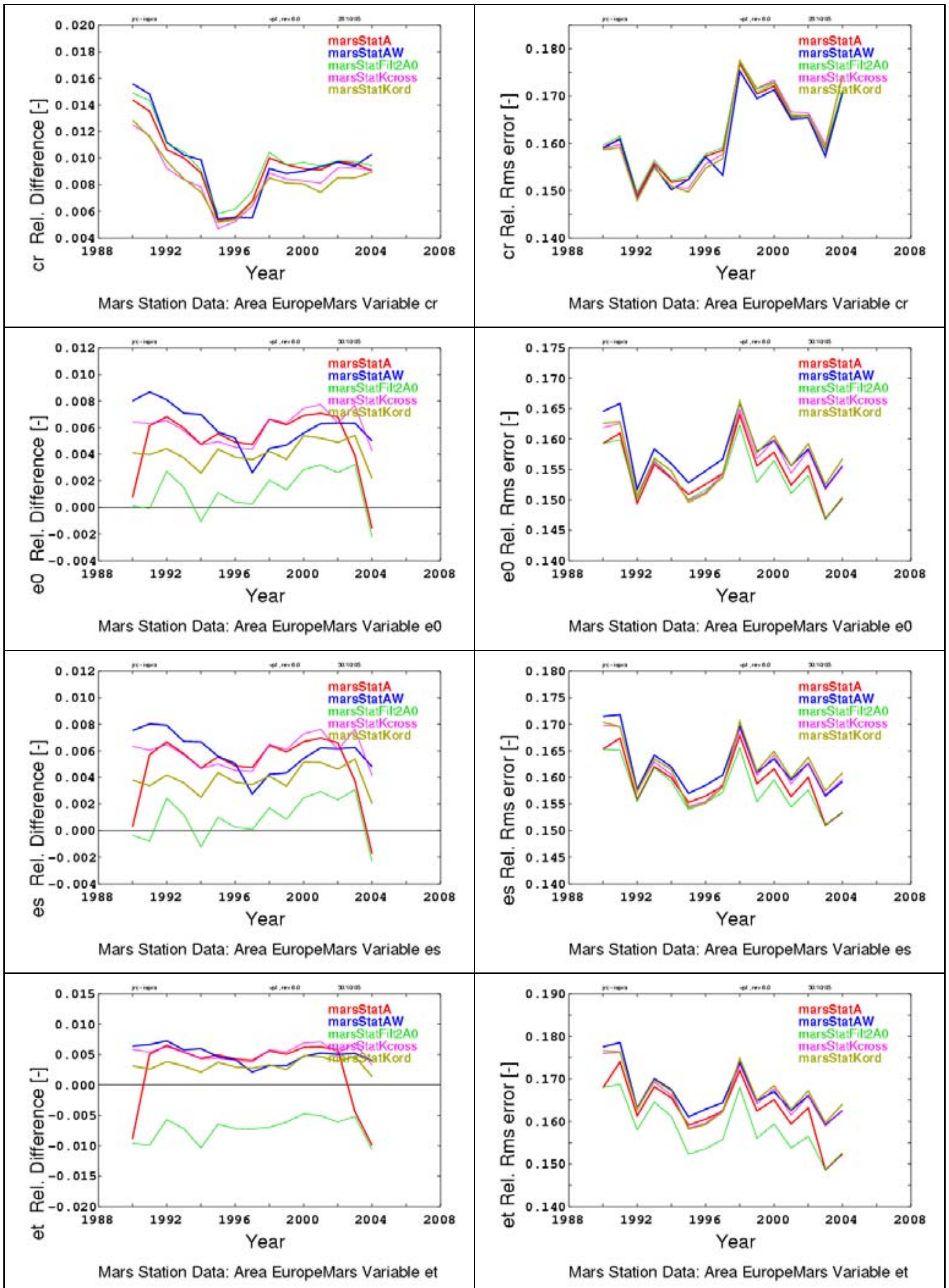


Fig. 5.1 – Mars Station data plots: cross validation of selected interpolation methods

In Fig. 5.1 and Tab. 9 are summarized the results obtained on the area named marsStation for the period 1990-2004. The *RelDiff* quantities are in the order of ± 1 %, with the exception of precipitation (± 4%) and wind speed (10÷15%). The RelEmse quantities are in the order of 10÷25 %, with the exception of precipitation (60÷80%) and wind speed (40÷50%).

		RelDiff									RelRmse			
	A	AW	Filt2A0	Kcross	Kord		Weight			A	AW	Filt2A0	Kcross	Kord
pr	4.3	3.9	3.0	3.8	2.1		6.0		pr	68.1	67.0	66.7	69.4	68.2
tn	0.9	0.3	1.8	1.1	0.6		3.0		tn	20.4	18.9	20.7	20.8	21.0
tx	1.0	0.6	1.8	0.9	0.8		3.0		tx	11.4	10.4	11.6	11.4	11.6
pd	0.9	0.8	0.9	0.9	0.7		1.0		pd	8.6	7.8	8.6	8.6	8.8
ws	12.7	11.5	12.8	12.7	11.8		1.0		ws	43.6	41.9	43.6	45.4	44.9
cr	0.9	1.0	1.0	0.9	0.8		1.0		cr	16.2	16.1	16.2	16.1	16.1
e0	0.5	0.6	0.1	0.6	0.4		1.0		e0	15.5	15.8	15.4	15.6	15.7
es	0.5	0.6	0.1	0.6	0.4		1.0		es	15.9	16.2	15.8	16.1	16.2
et	0.3	0.5	-0.7	0.5	0.3		1.0		et	16.3	16.7	15.8	16.6	16.6
Mean	2.6	2.3	2.5	2.5	1.8				Mean	34.5	33.6	34.0	35.1	34.7

Tab. 9 – *RelDiff* and *RelRmse* mean values in the year 1990-2004.
To the column with A is associated the data set marsStatA, to AW is associated marsStatAW an so on. See Table 8 for the interpolation parameters used.
For each row, with colour (pink) is identified the run with the worst quantities, while with colour (green) is identified the best one.
The *RelDiff* and RelRmese **mean** are the weighted average of the absolute values for each run, where the weight of the precipitation equals the total weight of min. and max temperature; it equals also the total weights of potential evaporations together with the other variables used in case of the correction in respect to the altitude difference is requested. The "mean" variable has been introduced in order to summarize in a unique value the results on the cross validation of the individual meteo variables.

Looking to the *RelRmse* mean (see Fig. 5.2) the runs that seem the best one are in order of preference are marsStatAW, marsStatFilt2A0, marsStatA, marsStatKord and marsStatKcross while in respect to the *RelDiff* mean are marsStatKord, marsStatAW, marsStatFitKcross, marsStatFilt2A0 and marsStatA.

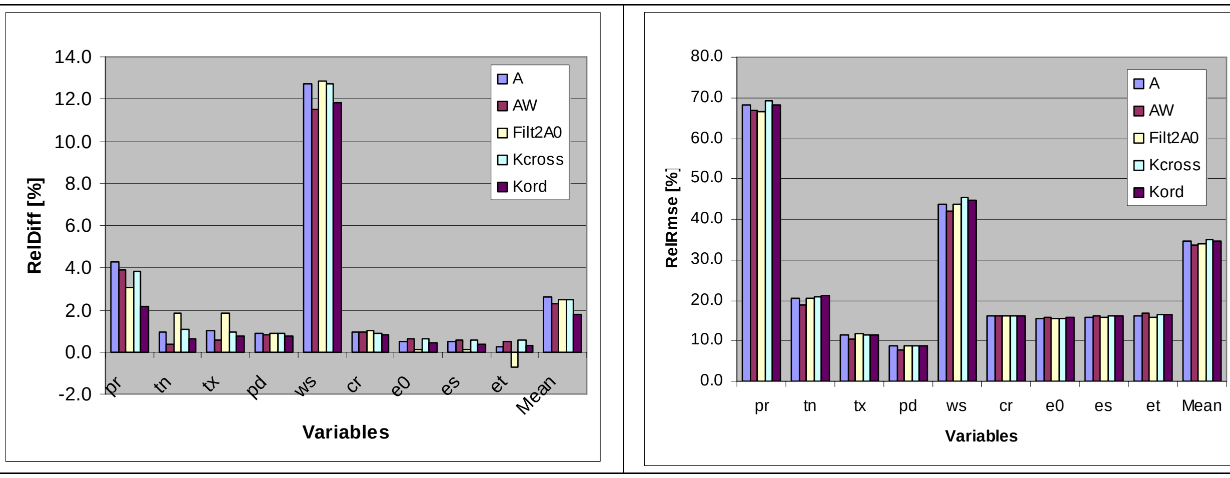


Fig. 5.2 - *RelDiff* and *RelRmse* mean values in the year 1990-2004 (from Tab. 9).

Summary and Conclusions

Different interpolation methods, together with filtering and weighting processes have been cross validated by using the Mars Station and High resolution rainfall data sets.

The interpolation results of the Mars Grid data and mars Station data (MARS-STAT data base) have been compared with the High Resolution data.

Indications on the best interpolation methods, together with filtering and weighting processes, have been applied on the full set of the meteo variables used as input by LISFLOOD (see Tab. 8)

Results

1) Any interpolation method used to interpolate the Mars Station data for EFAS hydrological simulations is better then the method used by CGMS to generate the Mars Grid data. In case of missing value at a station, the interpolation method seems better of using reference values obtained from historical data-sets.
2) For historical analysis and hydrological simulations, the interpolation parameters used to generate the data set marsStatAW (Inverse Distance method, *Nmax=15, Idp=1* and annual weight maps) are recommended
3) For setting up initial conditions of hydrological forecasts, the interpolation parameters used to generate the data set marsStatFilt2A0 (Inverse Distance method, *Nmax=15, Idp=1, filterFactor=2*) are recommended
4) If a standard error estimation is requested, the interpolation parameters used to generate the data set marsStatKord (Ordinary Kriging method, *Nmax=15, no Cleaning*) is recommended.

References

[1] van der Goot, E. (1997) 'Technical description of interpolation and processing of meteorological data in CGMS', JRC internal report
[2] P. A. Burrough, R.A. McDonnel (1998), 'Principles of Geographical Information Systems', Oxford University Press.
[3] G.Franchello (2005), 'station2map software', JRC internal report
[4] I. Clark (2005),' Practical geostatistics 2000', William V Harper, Otterbein College, Westerville OH, USA

Appendix 1 – Equations for Data Analysis

The Relative Root Mean Square Error (*RelRmse*) and the Relative Difference (*RelDiff*) between a reference field and a source field, are computed as follows:

$$\mathrm{RelDiff}_{\mathrm{C}} = \frac{\sum_{t=1}^{T}\left(r_{t,c}^{\mathrm{Sour}} - r_{t,c}^{\mathrm{Ref}}\right)}{\sum_{t=1}^{T}\left|r_{t,c}^{\mathrm{Ref}}\right|} \qquad \mathrm{RelRmse}_{\mathrm{C}} = \sqrt{\frac{\sum_{t=1}^{T}\left(r_{t,c}^{\mathrm{Sour}} - r_{t,c}^{\mathrm{Ref}}\right)^2}{\sum_{t=1}^{T}\left(r_{t,c}^{\mathrm{Ref}}\right)^2}}$$

where $r_{t,c}$ is the daily precipitation in the cells under analysis and T is the time period (days).

For *RelRmse* higher of 1 the evaluated error is higher of the observed rainfall, i.e., the error is higher of a source field with 0 values, which lead to *RelRmse = 1*. Note that the previous formulas are similar to the Nash coefficients

$$\mathrm{Nash}_{\mathrm{C}} = 1 - \frac{\sum_{t=1}^{T}\left|r_{t,c}^{\mathrm{Sour}} - r_{t,c}^{\mathrm{Ref}}\right|^{J}}{\sum_{t=1}^{T}\left|r_{t,c}^{\mathrm{Ref}} - r_{t,c}^{Banch}\right|^{J}}$$

In the used equations the coefficient *J* is 1 and 2 respectively and $r_{t,c}^{Banch}$ is set to 0. Note that in the *RelDiff* equation are not evaluated the absolute values like for the Nash coefficient because an evaluation on the differences between the total amounts of the quantities is wanted.

The map values are then summarized in a unique value by doing a weighted average of the cell values

$$\mathrm{RelRmse} = \frac{\sum_{\mathrm{c}=1}^{C}\mathrm{RelRmse}_{\mathrm{C}} \cdot Ndat_c}{\sum_{\mathrm{c}=1}^{C} Ndat_c} \qquad \mathrm{RelDif} = \frac{\sum_{\mathrm{c}=1}^{C}\mathrm{RelDif}_{\mathrm{c}} \cdot Ndat_c}{\sum_{\mathrm{c}=1}^{C} Ndat_c}$$

where *C* is the no. of cell in the area under analysis and $Ndat_c$ is the no. of days with precipitation values in the cells under analysis.

European Commission
EUR 22126 EN – DG Joint Research Centre, Institute for Environment and Sustainability

Luxembourg: Office for Official Publications of the European Communities
2005 – 52 pp. – 21 x 29.5 cm
Scientific and Technical Research series

Abstract

The observed meteorological inputs for the European Flood Alert System (EFAS) have been analysed. Observed sources are: 1) Grid and Station data collected from the MARS-STAT data base, 2) high resolution rainfall and mean temperature data covering Germany and upper Danube (HR). Several interpolation methods have been calibrated and cross validated. The comparisons of the different interpolated data sources are discussed and the "best" interpolation method and parameters are suggested. The software station2map is used to cross validate the different interpolation methods.

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.